Interdisciplinary Dialogues

The Major Risks of Generative AI

Guest: Yoshua Bengio

Speakers:
Caroline Lequesne
Hugo Loiseau
Jocelyn Maclure
Juliette Powell
Sonja Solomun

April 2024

obvia IVADO Mila

About the series *Interdisciplinary Dialogues*

In an exciting series of Interdisciplinary Dialogues on the societal impacts of AI, we invite a guest speaker and panellists from the fields of science and engineering, health and humanities and social sciences to discuss the advances, challenges and opportunities raised by AI. The first dialogue in this series began with Yoshua Bengio, who, concerned about developments in generative AI and the major risks they pose for society, initiated the organization of a conference on the subject.

The event took place on August 14, 2023 in Montreal, and was aimed at initiating collective, interdisciplinary reflection on the issues and risks posed by recent developments in AI. The conference took the form of a panel, moderated by Juliette Powell, to which seven specialists were invited who cover a variety of disciplines, including: computer science (Yoshua Bengio and Golnoosh Farnadi), law (Caroline Lequesne and Claire Boine), philosophy (Jocelyn Maclure), communication (Sonja Solomun) and political science (Hugo Loiseau). This document is the result of this first interdisciplinary dialogue on the societal impacts of AI. The speakers were invited to respond concisely, in the language of their choice, to questions raised during the event[1].

Immerse yourself in reading these fascinating conversations, presented in a Q&A format that transcends disciplinary boundaries. The aim of these dialogues is to offer a critical and diverse perspective on the impact of AI on our ever-changing world.

This Interdisciplinary Dialogues series is organized jointly by Obvia, IVADO and Mila

Obvia–The International Observatory on the Societal Impacts of AI and Digital Technologies is an open research network that brings together the expertise of over 260 researchers. Through critical interrogation, Obvia's mission is to identify the societal issues of AI and digital technologies, and to contribute to solutions that place living beings and the biosphere at the center of their development and use cycle. Obvia's research community, in collaboration with civil society, public actors, industry and developers, generates open knowledge and supports individual and collective capabilities.

IVADO is an interdisciplinary, cross-sectoral research, training and knowledge mobilization consortium whose mission is to develop and promote a robust, reasoning and responsible AI. Led by Université de Montréal with four university partners (Polytechnique Montréal, HEC Montréal, Université Laval and McGill University), IVADO brings together research centers, government bodies and industry members to co-build ambitious cross-sectoral initiatives with the goal of fostering a paradigm shift for AI and its adoption.

MILA – Today, the Mila community boasts the largest concentration of deep learning academic researchers globally. The institute is recognized for its expertise and significant contributions in areas such as modeling language, machine translation, object recognition and generative models. Since its inception, Mila focuses its mission on core research areas such as health, environment and climate change, and AI ethics. Mila extends its expertise and leadership in AI to deliver advances that will benefit all of society. Research carried out at Mila is conducted using an Open Science approach to promote collaboration and foster knowledge transfer.

1 Please note that Golnoosh Farnadi and Claire Boine were not available for this exercise. We did, however, benefit from their expertise and viewpoints during the August 14th panel at Mila.

Scientific Direction

Lyse Langlois
CEO, Obvia

David Hartell
Strategic Advisor, IVADO

Virginie Portes
Director of Research Support, IVADO

Support and coordination

Félix-Arnaud Morin-Bertrand
Research Professional, Obvia

Contributors

Yoshua Bengio
Scientific Director of Mila and IVADO and Full Professor at the Department of Computer Science and Operations Research of Université de Montréal

Caroline Lequesne
Associate Professor in Public Law at Université Côte d'Azur

Hugo Loiseau
Full Professor at the School of Applied Politics of Université de Sherbrooke

Jocelyn Maclure
Full Professor at the Department of Philosophy of McGill University and Stephen A. Jarislowsky Chair in Human Nature and Technology[2]

Juliette Powell
Entrepreneur, consultant and author in the field of technology and AI

Sonja Solomun
Deputy Director of the Centre for Media, Technology and Democracy and Ph.D candidate in Communication Studies at McGill University

Produced with the financial support of Fonds de recherche du Québec

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

ISBN: 978-2-925138-38-9
DOI: 10.61737/XSGM9843

2 Jocelyn Maclure was also president of the Commission de l'éthique en science et en technologie du Québec (CEST) from 2017 to February 2024.

Introduction by Lyse Langlois

"Consciousness without science and science without consciousness are both mutilated and mutilating."

Edgar Morin (1982/2017) [Our translation]

This inaugural edition of *Interdisciplinary Dialogues* kicks off with a look at the societal impacts of AI as examined by Yoshua Bengio, an international, award-winning researcher in *deep learning*.[3] In addition to the international recognition he has achieved in his research field, he is also a socially committed researcher who is as concerned about the benefits as he is about the risks that technological transformations can have on societies[4]. For Professor Bengio, the growing capabilities of AI, thanks in no small part to deep learning, suggest that it could reach levels of human intelligence over the next two decades. This potential capacity raises major risks and threats to democracy. Indeed, it is not a stretch to say that these AI-associated risks and misuses piggyback on the efficient accelerators that have become social networks. For instance, the spread of fake news and deep fake techniques are undermining public confidence. Increasingly, we are witnessing the polarization and fragmentation of societies attributed to these phenomena.

Will technology surpass human capabilities? Will technology replace humans within twenty years? This type of questioning, which is yet a further issue related to the main topics of discussion, raises major fears. One undeniable fact is that technology has already surpassed humans in certain capacities: memory, big data set analysis and processing, etc.; that is, technology not only performs certain tasks efficiently, but it also offers us a personalized, useful and uninterrupted relationship by responding to needs that sometimes haven't yet even been expressed. This is what some call the establishment of AI-based power of expertise in the conquest of human behavior. But what are we actually trying to achieve by striving to surpass certain human capabilities? Is the intent to reduce human autonomy and control human activity? What is the ultimate goal of these emerging, market-oriented technologies? For when it comes to sensitivity, free will, empathy and imagination, there is still a gap that we have yet to bridge. Does this gap represent the risk of extinction of human characteristics? Our contributors on this project don't provide us with direct answers to these relatively speculative questions. No one can. Instead, what these discussions have highlighted is the importance of using the right terms, of better defining notions of risk and everything that shapes our representations of technology while, at the same time, exploring in greater detail modes of rationality that are proposed to us.

There's no denying that technology is a major phenomenon with infinite social, economic and civilizational implications. Despite this, we still have the opportunity to regain control over the space we wish to accord it in our collective imagination and over the guidelines and regulations we wish to impose on it. After all, it is up to all of us as a society to get involved in these debates and to not shy away from teleological and normative questioning in the face of the inevitability of technological advances. The aim of this series is to open up a dialogue while presenting differing points of view on current issues.

"Human sciences are not conscious of the physical and biological characteristics of human phenomena. The natural sciences are not conscious of their place in a culture, a society, a history. The sciences are not conscious of their role in society. The sciences are not conscious of the occult principles that drive their elucidations. The sciences are not conscious that they lack consciousness. But everywhere is the need for a science with consciousness. It's time to become conscious of the complexity of all reality - physical, biological, human, social, political - and the reality of complexity. It's time to be conscious that a science devoid of reflection and a purely speculative philosophy are insufficient. Consciousness without science and science without consciousness are mutilated and mutilating"

Edgar Morin (1982/2017) [Our translation]

3 Together with Geoffrey Hinton and Yann LeCun, he was winner of the 2018 Turing Award.

4 From the outset, he has highlighted the importance of providing an ethical framework for AI through his support for the Montreal Declaration (2018), as well as by recognizing the importance of creating, as a co-investigator, the International Observatory on the Societal Impacts of AI and Digital (Obvia) within the Quebec ecosystem. He is also involved with the United Nations as an advisor to the Secretary-General on the impact of technology on society.

Reflection by Yoshua Bengio

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.

1. 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.

2. 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.

3 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.

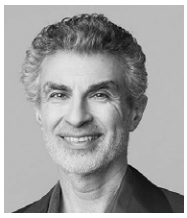
The magnitude of these risks is so considerable that we should mobilize our best minds and ensure major investments in these efforts, on par with past efforts such as the space program or nuclear technologies - in order to fully reap the economic and social benefits of AI, while protecting societies,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

Q

What do you think are the main risks to society posed by AI developments? How can we avoid them?

Yoshua Bengio[5]



The risks that worry me the most for the survival of democracy and humanity can be grouped into these three categories: (1) intentionally dangerous uses of AI, (2) excessive concentration of power that threatens democracy, human rights and geopolitical stability, and (3) loss of control of AIs that are sufficiently proficient in achieving their goals that they endanger society and humanity. The time horizons of these dangers overlap, but I've arranged them in chronological order of increasing anticipation. The uncertainty surrounding their respective scenarios and the magnitude of the expected negative impacts are also increasing, with the exception of the already documented human rights violations included in category 2, which could be better mitigated if the people and organizations currently empowered to make decisions about the use of AI were more representative.

Category (1) includes the use of generative AI for disinformation, fraud and the development of virtual/cyber attacks, chemical, or biological weapons. The worst-case scenario for (2) would be if a person, company or government took advantage of a significant advance in AI to seize excessive power, starting with economic, then political, then military domination. The simplest worst-case scenario for (3) is if, as some people are already expressing a desire to do, humans give a powerful AI the central goal of self-preservation, just like a human, and therefore the goal of doing everything possible to ensure that it cannot be stopped, which would undoubtedly lead to a conflict with an uncertain outcome if these AIs are more competent than us in enough areas to be dangerous.

5 All answers by Yoshua Bengio, Caroline Lequesne, Hugo Loiseau and Jocelyn Maclure were originally written in French and translated to English. The original versions of their answers are available in the French version of the document.

Caroline Lequesne[6]

The debate on the risks of AI brings with it considerations on the meaning of public action, its priorities and temporalities. While the darkest scenarios cannot be ignored, it is important to consider them in their tangible reality: they are the risks of tomorrow and represent one of the technological paths that are certainly possible - even probable - but not yet proven. We must guard against these risks - and anticipate them - but we can't devote all our political energies to this task. We see this as a scientific imperative, but also a democratic one. Recent debates on the subject have highlighted the harmful effects of such an approach: the sacrifice of the present in the name of a hypothetical future. The attacks on fundamental freedoms and their consequences for individuals and societies are numerous, real and well known. From the mobilization of AI to manipulate public opinion, to the pornographic deepfakes that disrupt the intimacy of our lives, there are many areas that require urgent action from our leaders. Failure to do so will have irreparable consequences for democracies and individuals alike. It would send a double signal: one of ignorance, especially towards the economically and socially vulnerable, who are the first victims of these systems, and one of *laissez-faire* towards the companies and administrations that deploy them. It is therefore essential to build solid systems of responsibility and accountability, capable of integrating all risks, while respecting freedoms, with graduated responses. In this respect, the approach adopted by the European legislator seems promising, as it moves away from contingency to propose present and future guidelines.

Sonja Solomun

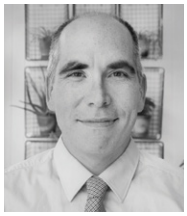
One of the central societal risks posed by emerging AI developments is the concentration and imbalance of power between those making decisions around how AI systems are built, deployed and governed, that is – those who benefit from AI – and those they directly impact. This imbalance is further exacerbated by the centralization of this power to a handful of companies who are currently making decisions that affect billions of people around the world. This power differential gets to the broader risks AI systems pose to democracy, by which I mean two key things: 1) the ability to preserve and act in the public interest and 2) the ability to affect decision making. AI developments need to be intentionally driven toward socially beneficial outcomes that safeguard our environment and the public interest over technological development and profit. Reciprocally, we need to prohibit AI where the harms to environment and public life are simply too great, that is, where surveillance predicates access to essential services like housing or education, and where the protection of human rights is jeopardized.

Finally, and perhaps most importantly, we sometimes take data for granted when we discuss the societal risks of AI, especially regarding generative and emerging AI systems. Since these systems are trained on massive datasets, we have to bear in mind the underlying data and assumptions about that data that AI systems make. AI systems do not simply reflect existing biases or social inequalities, they actively reproduce them as countless studies have shown. Here, work by leading scholars such as Simone Browne, Wendy H.K Chun, Safiya Noble, Ruha Benjamin and Timnit Gebru (to name only a few) is incredibly instructive in not only understanding how the very technical defaults and ground truths of AI are rooted in the past through longer systems of inequality, but crucially, how we can also mobilize present discriminatory outcomes to diagnose a future that we do *not* want. Climate change models are a generative example here – when global models identify the most likely future based on current and past data, we do not alter the model, we try to change our actions to avoid the "prediction" (Chun, 2015).

6 See page note 5.

Ultimately, the concentration of power we see today is not inevitable. Avoiding the democratic risks and imagining new engagements with AI that are not rooted in hierarchies of power requires us to think carefully and collectively about where in society such systems even belong and the decisions we make using them. Only then can we get to the "how" questions, including how their risks can be mitigated, how these systems should be built, who should be involved, how they should be overseen and by whom. Democracy crumbles not when we are unable to detect "truth" from manipulated or synthetic content but when we lose the ability to debate and make decisions about our collective future.

Jocelyn Maclure[7]



It is useful to distinguish between the risks posed by AI systems as they currently exist and those created by a possible general AI (AGI) whose cognitive capabilities would be equal to or superior to those of humans. While I believe that the risks associated with current AI systems based on various machine learning algorithms are significant and urgently require the deployment of a robust legal and regulatory framework, the actual recent progress in AI does not suggest that the emergence of a radically autonomous artificial superintelligence too cognitively advanced to be controlled by humanity is imminent, or even likely. Supporting this position would require an argument, of course, but it's crucial that those concerned about the implications of AI know that we can both be deeply concerned about the use of AI systems being developed in academic and industrial laboratories without thinking that we have good reason to believe that inorganic entities with no will of their own will seek -and be able - to dominate human civilization.

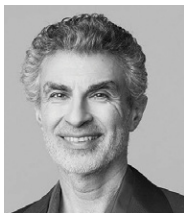
Among the risks posed by AI systems as they exist (and widely discussed in the burgeoning field of AI ethics), we can distinguish those inherent in their use to automate decision-making processes or other human practices (autonomous vehicles, content production, social robots, etc.) from those arising from the malicious use of AI by humans. On the one hand, the use of AI may lead, for example, to forms of discrimination prohibited by law, or to violations of the right to privacy in the absence of any malicious intent. On the other hand, ill-intentioned humans could use generative AI to misinform and pollute the public sphere. To mitigate these risks and to hold all stakeholders involved in the development and deployment of AI accountable, a binding regulatory framework must be implemented.

7 See page note 5.

Q2

What can be done to ensure that the development of AI benefits everyone, and to avoid the concentration of power?

Yoshua Bengio



The severity of this problem will increase as AI capabilities advance. One might doubt that market mechanisms will prevent the concentration of power resulting from very powerful AIs, since a company's goal is to establish as much economic dominance as possible, and money buys power. Competition between companies can counterbalance this, but the enormous computational costs associated with cutting-edge AI and the highly specialized talent required are already leading to a concentration of power in a very small number of companies. It is therefore important that nonprofit organizations whose mission is the public good are able to (a) assess the risks associated with other organizations' AI, for example, through mandatory regulatory audits, and (b) offer an alternative to AI companies whose goals are, on the one hand, to protect the public and, on the other hand, to apply AI for the public good, such as the UN's SDGs, under inclusive and democratically established governance that gives everyone a voice, including defending the interests of countries with no access to such power.

Sonja Solomun

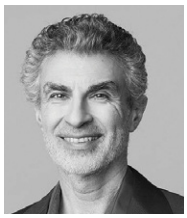
This is a crucial question because it presupposes an important distinction – there is a difference between making sure "harmful" outcomes do not happen (which many regulatory, academic and civil society frameworks take as their basis) and ensuring "beneficial" things *do* happen, especially for those who stand to be most adversely impacted by AI. It is no longer sufficient to simply minimize the harms of AI since what is left is never a neutral "tool" but rather, a dynamic and evolving socio-technical system that either implicitly or explicitly makes political assertions depending on where and how it is used in society (in other words, it would not address the "misalignment problem" Yoshua Bengio spoke of earlier). Moving beyond harm reduction approaches also allows us to ask collective questions about whether a particular use case is even justified, or whether alternate frameworks exist.

One way to direct AI development in the public interest and to ensure we are evaluating these systems based on social impact is through independent regulation that is intentional about empowering outcomes toward the public good. AI governance needs collective and ongoing input from a plurality of voices, especially from affected communities (through possible review committees, standard-setting and other mechanisms, including enforcement). Regulatory AI frameworks also need transparency about both model and data documentation and perhaps most importantly, robust accountability mechanisms including a liability regime, strong data rights with clearly defined categories and prohibitions of high risk, as well as frameworks which address AI deployment and risk beyond individual rights and harms to include systemic and environmental risks.

Q3

How can we preserve democratic values in the face of the malicious use of AI?

Yoshua Bengio



The rise of AI raises several concerns for democratic values and institutions. On the one hand, the malicious use of AI could lead to increased state surveillance, also supported by AI, with the stated aim of protecting the population. However, it could also justify and enable governments with authoritarian tendencies to increase and centralize their power through AI, threatening democracy. A malicious use of AI by other states targeting our democracies could lead to an AI-based arms race, a dangerous situation both for geopolitical stability and in terms of mistakes that could lead to AI losing control. It therefore seems essential to me to establish a democratic and international governance of AI that would prohibit its development for the purposes of military or political domination. Organizations developing cutting-edge AI should first and foremost serve the public good, democratic institutions, the UN's Sustainable Development Goals, and the defense against malicious uses of AI for such power.

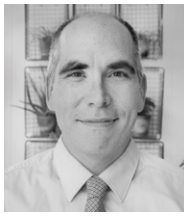
Caroline Lequesne

The question itself is a challenge because it confronts two inherently irreconcilable dynamics: to destroy through AI or to build together. The key is not to sacrifice the latter for the former; not to let fear paralyze virtuous processes of sharing, debate, and fundamental rights. We are aware that democracy will not provide all the answers we hope for. However, this state of affairs should not lead us to abandon the democratization of AI. The latter is based on two fundamental elements: first, a framework that respects fundamental freedoms; and second, the involvement of citizens.

The first element is reminiscent of the debates of the 2000s on the expansion of the market sphere and the financialization of society: did they herald the end of democracy? The work of many economists, most notably Jean-Paul Fitoussi, has shown that the market cannot survive without democracy. The same is true of technological innovation: AI law is not intended to be a law of exceptionality and secrecy but must be integrated into a regime of freedom. Moreover, it seems essential that this right should remain the right of all. To this end, the citizen, who has so far been sidelined in the name of technical considerations - or imperative economic and security issues - has been largely excluded from the debate. And yet, how can we build a society

without the participation of the vast majority? In this respect, it seems essential to institutionalize a technological public space alongside laboratories. Sociological studies on nuclear energy have demonstrated the virtuous complementarity between scientists and citizens. This technological citizenship remains to be conceived and built.

Jocelyn Maclure[8]



This answer will displease right and left-libertarians alike, but the only realistic approach in the world we live in is that of robust state interventionism: rulemaking, regulation, and redistribution. Existing sector-specific laws have been or are currently being revised to remain relevant and effective in the AI and digital age (privacy laws, competition laws, consumer protection laws, copyright laws, etc.). Laws specific to the development and use of AI must be enacted (AI ACT, LIAD, etc.).

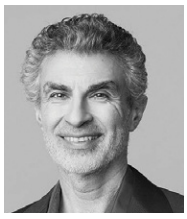
Various means of taxing the revenues of major web companies must be deployed to support ecosystems severely affected by the digital revolution (media, artists, etc.) and, more generally, to enable states to fulfill their responsibilities in terms of redistributing wealth and reducing digital divides. We can of course hope for greater democratization of businesses and workplaces, but this transformation will not happen spontaneously; it will require legislative changes.

8 Jocelyn Maclure noted that he was answering questions 3 and 4 simultaneously.

Q4

What measures need to be put in place to protect society in terms of cybersecurity (regarding the malicious use of AI)?

Yoshua Bengio



The most important measure would be to require that very large language models capable of programming above a certain threshold be registered with the government, not be distributed over the Internet, not be accessible over the Internet, or not be sufficiently protected from cyberattacks to allow their code and settings to be copied. In addition, care should be taken to remove from training data any code used for cyberattacks. Second, international agreements should be made to ensure that such measures are applied worldwide. In addition, investment should be made in an international network of non-profit, publicly funded laboratories whose goal would be to implement defensive measures (including the monitoring of otherwise silent cyberattacks) using secure AI (which in itself also requires investment in research).

Hugo Loiseau[9]

The geopolitics of cybersecurity is unique in that it is a field that increases its attack surface in proportion to the development of its arsenal. The same will apply to AI. The question is how to control that arsenal. Studies in cybersecurity tell us that non-proliferation of cyberweapons is hardly working at the moment because states are in a logic of power supremacy in cyberspace, which is a rather classic security dilemma. Moreover, states act according to a principle of internal balancing (Waltz, 1979), whereby an innovation that gives one state a power advantage will be copied by other states. Applying this logic to AI, the most powerful states will want to control what other states do, while allowing themselves the freedom to develop their own AI-based arsenals according to their own interests. In this respect, Joseph Nye's (2011) essay on what cybersecurity can learn from nuclear power as a dual-use technology is highly applicable to AI. In this sense, the governance needs for AI are tremendous (see next question).

9 See page note 5.

Juliette Powell

Just as mechanical engineers and electrical engineers are held accountable by their profession, computer and systems engineers need to be trained and retrained/certified and recertified in responsible AI so that if/when humans are harmed by the deployment and/or use of these technologies, computer and systems engineers are held responsible for their creations, including when they leave vulnerabilities that can be hacked. Traditionally "operators" are held accountable as, for example, was the case of the 2018 Uber self-driving car accident in Tempe Arizona that killed a woman crossing the street with a bicycle.

In terms of global accords, they will need to be aggressively enforced and it isn't clear how they can be while also prioritizing innovation and competitive advantage.

From that perspective, a global network of non-profit labs may be necessary to provide independent evidence of potential harms to humans. These new global institutions would require equal representation, including a large number of people from outside the field of AI. These cross-disciplinary teams who understand sociology and psychology, philosophy, anthropology, and law would be key to dealing with these very large questions on behalf of humanity, in addition to the other logics of power around AI: corporate, engineering, government, and academia.

Q5

At what level should we establish AI governance (local, regional, national or international)?

Hugo Loiseau

In the field of AI, international cooperation and international law are lagging behind the risks and threats associated with AI. As a result, there is an enormous need for governance at all levels. Local and municipal, especially for the social acceptability of AI and its infrastructures, which consume energy and water for cooling. Regional, for all cross-border governance issues (trade, value chains, flow management, etc.). National, for research and development policies, but also for the framework of AI in a national legal order that guides local, regional and international levels. Finally, at the international level, for the development of an international AI law aimed at regulating, as far as possible, research centers and companies (digital or otherwise). At the international level, the model to follow, in my opinion, would be that of the nuclear non-proliferation regime. While far from perfect, this international regime has allowed the emergence of international norms implemented by the International Atomic Energy Agency (IAEA) through the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). This UN agency promotes the peaceful use of the atom, while limiting its military applications.

Caroline Lequesne

The question calls for a distinction between *what is* and *what ought to be*, between the realm of well-founded hope and the realm of pragmatic action. While the two levels of response are not irreconcilable - and feed each other - the former invites us to free ourselves from the latter in order to be "ambitious". Given the economic, democratic and civilizational issues raised, an international response would be advisable. An agreement with the broadest possible support, aimed at reducing competition and its effects in favor of a general, common and human interest, would be, if not a panacea, the symbol of a much-needed collective realization. However, the experience of the "thirty globals" on major planetary challenges bears witness to the weakness and limitations of international law. At the same time, national initiatives appear modest, if not inadequate. Therefore, from a pragmatic point of view, the regional level seems to be the most promising. In this respect, Europe shows the way: while compromises remain complex to build, the level of requirements is raised and the implementation of rules, within a large consumer market is likely to have a knock-on effect in other zones (where we can again speak of the "Brussels effect", theorized in corollary fields such as personal data protection).

Juliette Powell

From a global perspective, three companies: Microsoft, Google (Alphabet) and Amazon are vying to be the #1 cloud AI provider in the world. In an industry growing 70% or more each year, the ability of regulations to constrain behavior is difficult. Moreover, the use of LLMs may seem like a fairly small change in the way we operate but their impact could change the world.

As such, we should establish AI governance at every level, beginning at the level of the individual. As stakeholders in the AI ecosystem, we have the responsibility to use critical thinking when vetting and experimenting with emerging technologies. We also owe it to ourselves, at every level, to develop a calculus of intentional risk around AI. Risk functions can estimate extreme changes, but they are not random. For organizations and governments, gains or losses can usually be estimated – and with enough deep knowledge about a particular process or event, an estimate can be reasonably reliable. For example, banks use risk functions to evaluate credit-worthiness; a person with a good credit score is a better risk than someone with poor credit and few resources. To establish a calculus of intentional risk, a decision maker could begin by recognizing the significant potential consequences for each AI system venture. For each consequence, they could create a matrix, plotting the severity of impact against the likelihood of occurrence; Establish a risk value for each matrix cell, based on those matrices; Create scenarios for the combination of risks, which provide a solid sense of the range of outcomes; Make a decision based on those scenarios, establishing ways of rapidly recognizing when there are surprises or anomalies, so the venture can quickly shift direction.

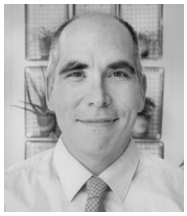
At the local, regional, and national level, corporations often have more power than governments to effect rapid change. Thus corporations and governments need to work with each other, as well as with the other logics of power around AI, academic, engineering and social justice logic, on an even playing field so that all logics are represented equally through any governance initiative.

Sonja Solomun

While I echo the call for governance on every level, robust national regulation is crucial, and should ideally be implemented through an independent regulator with strong oversight and enforcement mechanisms. One option follows the European Union's risk-based model, meaning use cases are regulated rather than a given technology "as such". While the risk-based approach is not without its faults, it is a starting point to frame the regulatory debate away from simply mitigating harms toward more proactive measures – before potential or demonstrable harm occurs. Risk assessments and independent audits (especially prior to data collection and deployment of AI systems) would be central accountability mechanisms to follow. Yet, minimizing the risks of AI systems is nevertheless a partial regulatory solution given the continuation of harms that will persist despite robust risk assessment, testing and mitigation. This is especially the case for systemic harms such as the energy cost of large language models for example. Beyond minimizing risks, AI regulation should be broadened beyond individual conceptualizations of risk, rights and safety to include systemic and environmental risks (Ada Lovelace Institute, 2023). As a starting point, regulation should be purposeful in driving AI development and use toward the public interest, which could be the task of a global governance body.

Relatedly, we urgently need greater international coordination of AI governance, especially since countries have distinct national laws and priorities. Democracies around the world need to be pulling their own distinct policy and regulatory levers toward a common framework of robust accountability and oversight of AI systems. This will require concretizing global models of AI governance beyond values and principles to set out formal commitments rooted in human rights frameworks. Crucially, a global AI governance body must have enforcement powers beyond simply advising on commitments. Global AI governance should also ideally work towards common parameters including environmental justice, and include strict specific prohibitions for unsafe AI systems or use in high stakes areas. Importantly, majority world and global south countries must be included in global AI governance such that the regulatory landscape is not led by European and North American values and markets.

Jocelyn Maclure

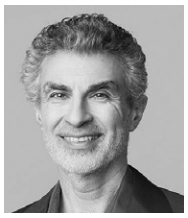


While international governance of AI systems is urgent and necessary, we must resist the temptation to think that it is pointless to regulate AI at other levels of political governance (urban, regional, national, continental). Democratic states are responsible for ensuring that fundamental rights are respected and to set the rules for the operation of institutions and sectors critical to the deployment of AI (health systems, electoral processes, the judiciary, transport, telecommunications, education, etc.). While the development of binding international law is highly desirable to promote a levelling of norms and standards governing the development and commercialization of AI systems, it would be tragic if states and local governments failed to develop robust AI governance mechanisms in the short term. Furthermore, although it is a non-binding text, it should be noted that UNESCO's Recommendation on the Ethics of Artificial Intelligence was adopted by the organization's 193 member states after extensive multilateral negotiations.

Q6

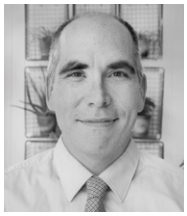
Are AI developments redefining the properties of human intelligence? Can we be "surpassed" by AI?

Yoshua Bengio



AI research teaches us that intelligence is diverse (an entity can be intellectually competent at some things and not at others) and can be quite different from our own (which ethology has also told us), even when human intelligence has been a major source of inspiration. This is both exciting, because we could learn a lot from other forms of intelligence, and worrying, because future AIs could interpret the world, and in particular the morality of actions, in a way that is very different from ours, and also potentially dangerous for humanity. Based on our knowledge of neuroscience and AI, there is no reason to believe that human intelligence is an unsurpassable pinnacle of intelligence, on the contrary. Again, this is exciting, but it's also worrisome if we don't proceed with extreme caution down the path of AI systems that are smarter than we are. However, algorithmic theory in computer science suggests that any intelligence will also be limited (and thus makes the concept of singularity in AI rather implausible), because many computational problems are unsolvable with limited computing power: so there is also an upper limit to intelligence.

Jocelyn Maclure



From its origins to the present, AI research helps us better understand the conditions of possibility, the strengths and limitations of animal, human, and non-human intelligence. I can easily understand why computer scientists and engineers speculate that AI programs will eventually make machines roughly as intelligent as humans (human-level, as Yoshua Bengio writes in the introductory text), or perhaps even more intelligent ("superintelligence"). The fact remains that animal intelligence is the fruit of a long biological evolution that has enabled certain living organisms to adapt to their environment and reproduce. It is quite plausible that some of the properties we associate with intelligence are not achievable on the basis of inorganic matter and in the absence of the biological cells that keep an organism alive. It's perfectly possible to be skeptical about the attainment of AGI while holding to an entirely naturalistic or materialistic view of the world and how it works.

Human cognition, moreover, includes not only the (real, but fallible) ability to reason, but also desires, emotions and intuitions. It stands at the intersection of affectivity, consciousness, will and reason-notions that are both essential to understanding human beings and difficult to pin down scientifically. Human cognition is a matter of both common sense and rationality.

Cognitive science is no longer exclusively dominated by behaviorism and its more sophisticated successors (functionalism, computationalism, etc.). The triple-E (embodied, embedded, enactive) theory of cognition has been one of the most influential in recent decades. This theory of cognition, inspired by German and French phenomenology, is sensitive to the role of the body (embodiment), of inclusion in an environment that is both natural and cultural, and of (inter)action within this environment. It's conceivable that giving a physical (robots) or digital (avatars) "body" to an AI that has to act in a physical or virtual environment won't bridge the ontological gap between artifact and living organism that I alluded to earlier. The fact that all these theories remain speculative for the moment does not make them equally plausible in light of what we know about human cognition.

Q7

How can we, as an academic community, better communicate on AI issues to inform public discourse?

Juliette Powell

Ultimately, governments, corporations, academics, social justice advocates, and engineers must not only establish internal and external governance that protects the majority of people on the planet, we must also communicate in such a way that non-experts can understand. The academic community is perfectly positioned to spark a global dialogue that allows more people on the planet to ask ourselves better questions about AI as a global community. For example, why are we developing AI? Is it at the service of the human race or of the technological race; Are we building AI for the arms race or for the human race? Another question to be considered by all AI stakeholders at every level: What are we NOT willing to do to make money; what are we NOT willing to do to gain competitive advantage; what are we NOT willing to do to gain global AI supremacy. From there, the cross-disciplinarian academic community must communicate where these lines are drawn internally and externally so we can all feel we have skin in the game and are held accountable by this transparent new global social contract. As Bengio points out, an international governance structure would ensure that human rights are at the center of any AI governance/regulation that would actively be enforced by seeking out, and restraining bad actors at every level. But for that to happen, the people of the world need to feel like they are part of the process without being felt like they will be left behind.

Hugo Loiseau

Academics will be able to better communicate and inform not only discourses, but also decisions about AI through a method called Responsible Research and Innovation (RRI), which is defined as "a transparent, interactive process in which societal actors and innovators engage with each other with respect to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (to enable the proper embedding of scientific and technological advances in our society)." (Von Schomberg, 2013). This means working with the theoretical and methodological tools at our disposal as researchers. We must work on this great interdisciplinarity and intersectorality (research, society, business, government). Implementation is certainly difficult and demanding, but seems possible to me, with the ultimate goal of developing trust between all these actors with regard to AI. This method has already been used advantageously, as demonstrated by the UK's Public Dialogue with Society, the work of OBVIA and INGSA, for example, on AI or other technological innovations.

Conclusion

The responses to the various questions provided a transversal and interdisciplinary perspective on the main risks posed by recent advances in AI, but also allowed us to reflect together on how to prevent them and ensure that the development of this technology benefits everyone. In this regard, the need for international governance of AI is recognized by all, although the importance of local, regional and national frameworks is also put forward by the contributors. As for the main issues identified, despite some elements of disagreement, it was possible to identify several points of convergence between the answers of the various contributors, particularly regarding issues of democracy and human rights.

Main issues summary:

Democracy and human rights

- Malicious uses, especially of generative AI, threatening the electoral process and public debate (production of deepfakes, creation of fake accounts on social media)
- Use of AI by states as an instrument of domination and control, both internally (within their populations) and externally (cyberattacks between states).
- Risks inherent to AI models, such as invasion of privacy and discriminatory effects of algorithms, which mainly affect the most people and groups who are the most marginalized and vulnerable groups.

It should be noted that many of the risks threatening democratic life and human rights are exacerbated by the increasing centralization of power, in the sense that the development of AI, while affecting all individuals, is (and will increasingly be) controlled by a small number of private companies.

Environmental costs and loss of control of AIS:

In addition to issues concerning democracy and human rights and freedoms, the environmental risks arising from the energy costs of AI models and the dangers associated with a possible loss of control of AI systems, or even artificial "super intelligence", are also raised. However, we note a point of disagreement regarding the plausibility of this loss of control of AI systems, mainly concerning the prioritization of risks arising from this scenario. This divergence of viewpoints seems to be explained, at least in part, by differing visions of the capabilities of artificial intelligence and of what constitutes human intelligence.

Finally, we would like to thank Yoshua Bengio, our first guest, and the other speakers for sharing their insights. This exchange of different, sometimes even conflicting, perspectives and viewpoints, which took place in a context of open and respectful dialogue, has helped to enlighten the debate and enrich our understanding of the societal implications of AI and how this technology is transforming our societies. This initiative is therefore intended to be a starting point, the first step towards greater collaboration and ongoing conversation between specialists in science and engineering, the humanities and social sciences, and health.

References

Ada Lovelace Institute. (2023). *People, risk and the unique requirements of AI: 18 recommendations to strengthen the EU AI Act*. https://www.adalovelaceinstitute.org/policy-briefing/eu-ai-act/

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

Browne, S. (2015). *Dark Matters: On the Surveillance of Blackness*. Durham: Duke University Press

Chun, W. H. K. (2015). On hypo-real models or global climate change: a challenge for the humanities. *Critical Inquiry*, 41(3), 675–703. https://doi.org/10.1086/680090

Chun, W. H. K. (2021). Discriminating data: Correlation, neighborhoods, and the new politics of recognition. MIT press.

Gebru, T. (2020). Race and Gender. Dans Dubber, M. D., Pasquale, F., et Das, S. (Eds.), The Oxford Handbook of Ethics of AI (Ser. Oxford handbooks). Oxford University Press.

Morin, E. (2017). Science avec conscience. Éditions Points. (original work published in 1982)

Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*. New York University Press

Nye, J. S. (2011). Nuclear lessons for cyber security?. *Strategic studies quarterly*, 5(4), 18-38.

Quantum city (United Kingdom). (2020). *Public Dialogue Report on Quantum Technologies*. https://www.quantumcity.org.uk/news/public-dialogue-report-quantum-technologies

Von Schomberg, R. (2013). A Vision of responsible Innovation. Dans Richard Owen M. Heintz and J. Bessant (eds.), *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, pp. 51-74. Wiley.

Waltz, K. N. (1979). *Theory of international politics*. McGraw-Hill

obvia IVADO Mila