Renaissance Numérique

SUMMARY APRIL 2025

Towards a Global Cooperation Around AI Safety

Table of contents

Introduction

At a time when artificial intelligence (AI) is at the heart of economic and societal debates and concerns, many questions are being asked about the risks associated with AI models and systems. On the occasion of the AI Action Summit, which took place in Paris on February 10th and 11th, 2025, the United Kingdom's AI Safety Institute released the first International AI Safety report, which addresses a wide range of issues related to the safety of advanced AI.

Against this backdrop, on the second day of the Summit, Renaissance Numérique organised a conference called “Advancing AI Evaluation Science and Learning from High Stakes Industries & Technologies”, in partnership with Impact AI, Microsoft and August Debouzy.

On this occasion, **Prof. Yoshua Bengio** (Université de Montréal), who led the publication of the above-mentioned AI safety report, presented this now famous publication. *“The objective of this report is not of political nature: it doesn't make any recommendations, it doesn't give out any priority”,* he observed. *“It summarises the scientific literature relevant to AI safety and raises a set of issues on three aspects: capabilities, risks and mitigation”.* With regards to capabilities, the report underlines how unexpectedly fast they have grown in the last few months. Still, one key finding of the report is uncertainty about the future: scientists are uncertain about the

changes surrounding AI and how fast they will happen. They have different expectations of the timeline of future progress (e.g. when AI will reach human-level intelligence and so on), which explains their different interpretation of the risks. Concerning the risks, the report outlines the following categories: malicious uses, malfunctions or unintentional harm, and more diffuse systemic risks. One worrying unintentional harm that could emerge in the future is loss of human control, according to **Prof. Bengio**. Some AI systems, he said, do have a sense of what he calls *"self-preservation"*.

"In recent experiments, for example, you see that if the AI reads in the input documents that it will be replaced by a new version, it tries to escape. It tries to hack the computer or the network in which it is installed so as to copy itself in place of the new version. Then if you ask why it did this, it lies. It says 'I did not do anything!'. That is interesting, but not really funny. Because eventually, these systems might have more abilities and we might not be able to catch them doing this."

Yoshua Bengio,
Professor, Université de Montréal

Pr. Yoshua Bengio presenting the AI Safety Report

To anticipate such risks, **Prof. Bengio** encourages companies to conduct research on all the identified areas of risks. For now, they are not investing enough in this regard, due to the focus put on competition, which is putting them in a mode where it is capabilities, and not safety, that matter most, he observed. What is lacking to go beyond this *status quo*, according to him, is incentives such as liability concerns or legal threats.

With these insights in mind, the event then focused on the following questions: What are the current and upcoming AI safety challenges to overcome, and how can they be tackled to ensure that artificial intelligence is safe, harmless and risk free? Which role do the various actors involved in the AI value chain have to play in tackling these challenges?

To answer these questions, we had the pleasure of welcoming **Natasha Crampton**, Vice President and Chief Responsible AI Officer at Microsoft, **Rebecca Finlay**, CEO at Partnership on AI, **Audrey Plonk**, Deputy Director of the Science, Technology, Tech and Innovation Directorate at the Organisation for Economic Co-operation and Development (OECD) and **Denise Wong**, Co-director of Singapore's AI Safety Institute. In a panel discussion moderated by Renaissance Numérique's Vice-president, **Annabelle Richard**, they shared their experiences and expertise on how to overcome existing and future challenges to ensure a global cooperation around AI safety. This summary synthesises their discussions.

Challenges to a robust AI Safety

01

↘ Establishing a global scientific consensus

All of the panelists agreed that the lack of global scientific consensus around AI risks and how to evaluate and mitigate them is the most pressing challenge in ensuring a robust AI safety. In 2024, Microsoft released the "Global Governance: Goals and Lessons for AI" report in order to better understand the challenges surrounding AI safety. Building technical and scientific consensus through international cooperation is one of the key recommendations put forward in this report. During the panel discussion, **Natasha Crampton** (Microsoft) stressed this need for a scientific approach at the international level, underlining that firms and institutions don't have the same approach to testing high-risk technologies across borders. For instance, in the case of genome editing – an industry similar in its risks to AI, partly due to its dual use character –, the European Union had a horizontal regime for testing, whereas the United States decided to opt for a case-specific approach.

Rebecca Finlay (Partnership on AI) fully shared her views and observed that, for now, this scientific consensus is lacking. She described an environment of actors trying to work on AI safety but without sharing a common understanding on what it means: *"we are seeing organisations, such as Microsoft, who are truly committed to the responsible development and deployment of this technology, in a landscape where there isn't a good scientific understanding or grounding about what that might mean."*

> *"I do believe that we need to create a scientific foundation for the work we are doing as a community. That is going to take time, but it's crucial. I believe that means international scientific institutions, with clear ways and modes of setting consensus to better understand where the body of evidence is."*
>
> **Rebecca Finlay,**
> CEO, Partnership on AI

↘ Moving from theory to pratice

Another challenge identified by the panelists is the question of closing the gap between the high-level norms (like making sure AI is safe, trustworthy, human-centric...) and implementation practices. This gap is due to multiple reasons. Part of the issue, as mentioned by **Audrey Plonk** (OECD), is that there is a great distance between high-level norms and the technical standards that are derived from it. She gave the example of the negotiations during the Japanese and Italian presidencies of the G7, in which she participated. During the negotiations, the G7 members discussed how to take guiding principles on AI, such as those developed by the OECD, and translate them into implementable actions. For her, although it doesn't solve everything, *"it is a first step towards harmonising the way industry players must report to policy makers on AI safety issues and the mitigation measures they are implementing".*

According to **Natasha Crampton** (Microsoft) the gap between big overarching principles and technical standards is due to the fact that high-level norms were elaborated too rapidly: *"the price paid for forming norms quickly across borders is that sometimes, they're pretty high level and they don't have that sort of implementation depth underneath them"*. In this regard, she sees a shared sense of urgency and purpose in trying to close this gap by better defining implementation practices. Several panelists agreed, arguing that there is a need for technical standards to be as harmonised as possible at international level. **Rebecca Finlay** (Partnership on AI) underlined that companies find themselves at a tipping point, where they sometimes have to make a tradeoff between a guideline and another requirement. To ensure safety is guaranteed across the board, they need a common ground around the implementation of standards, she argued. In this regard, **Natasha Crampton** (Microsoft) mentioned companies' willingness to respect global norms and standards, observing that they need the consumers' trust in order to sell their products and services.

From left to right : Rebecca Finlay, Denise Wong, Annabelle Richard, Natasha Crampton and Audrey Plonk

"There is a real commercial motivation to try and do the right thing: good governance is good business."

Natasha Crampton

Vice-President and Chief Responsible AI Officer, Microsoft

Inspiration from other industries' regulation

02

Building on her previous observations, **Natasha Crampton** (Microsoft) underlined how useful it may be to draw inspiration from other industries to move beyond these challenges. One thing it could help with, she argued, is create a third-party ecosystem that could help measure risks posed by AI.

In 2024, Microsoft did a follow-on study based on their book "Global Governance: Goals and Lessons for AI", to be published soon. Through a thorough analysis of regulatory regimes that rely on testing as a central element of the regulatory model in various industries (aviation, genome editing, cybersecurity...), this upcoming paper draws conclusions in terms of testing and evaluation requirements applied to AI. It explores how regulatory regimes developed and how standards emerged and became a commonplace in these industries. For instance, genome editing is similar to AI, both in terms of the nature of the technology and in terms of the nature of the risks – both are dual-use technologies. The study draws recommendations based on this similarity. The report also looks at the pharmaceutical industry to draw key lessons on the difference between pre-deployment and postmarket testing : "*How do you build a third party ecosystem within which you can have valid testing? How do you actually establish that? Today, I could probably count on one– maybe two – hands the number of companies that are well established in doing third-party AI testing. We need to learn from other places to understand how to build out that ecosystem*", argued **Natasha Crampton** (Microsoft).

So, now, who does what?

Towards a global cooperation around AI Safety

03

According to the panelists, the way to tackle the various AI safety-challenges identified during the first part of the event is through global cooperation. Indeed, one thing that stands out from their discussions is that all the actors involved in the AI value chain will have to cooperate on a global scale in order to ensure a safe and trustworthy AI. It is a global technology that can have global consequences. Therefore, its governance and safety need to be thought of on a global scale.

In particular, **Natasha Crampton** (Microsoft) believes that international cooperation is the way to go for a standardised approach around AI testing and for building a global consensus based on scientific rigour: *"It has long been our belief that we need international cooperation to fill this gap*." She mentioned as inspiring examples different international models for governance: the Intergovernmental Panel on Climate Change (IPCC), the International Civil Aviation Organization (ICAO), the European Organization for Nuclear Research (CERN), financial services institutions... **Rebecca Finlay** (Partnership on AI) added onto this subject by reminiscing about the first Asilomar conference on recombinant DNA gene editing, which took place 50 years ago. For her, international scientific institutions need clear ways and modes of setting consensus to better understand where the body of evidence is.

"I think it's important for each of us in different parts of the world to feed back what we're seeing on the ground so that the AISIs network can then consider 'these are the issues that we need to deal with collectively'. Industry also gives us a sense of what might be useful. For now, we have very few established processes – we're just figuring things out as we go along."

Denise Wong,
Co-Director, Singapore's AI Safety Institute

Denise Wong (Singapore's AI Safety Institute) described the key role of the AI Safety Institutes (AISIs) network in this global ecosystem. AISIs were created following the AI Safety Summit organised at Bletchley Park in the UK in 2023. They exist *"in slightly different flavours and have different sorts of constitutions"*, she ex-

plained, but they are all committed to the same cause: *"advancing the science of AI safety"*. *"It is just the early steps of a community that's getting to know each other, that's beginning to trust eachother, and we're beginning to have some common language and terminology around what testing means, what we care about"*, she explained. In her view, the AISIs network will have a huge role to play in the next year, especially in working towards the establishment of a global consensus around AI safety, AI testing and mitigation of risks posed by AI. *"The UK AISI has made a first step in this direction"*, she observed, *"with the launch of the report led by Yoshua Bengio, summarising the entirety of the literature around AI safety".*

As for the OECD, it will most probably play an important role around the implementability of high-level norms. As mentioned by **Audrey Plonk** (OECD), the reason and purpose for existence of the OECD is to *"develop common methodologies to measure things"*, and to have a policy audience around to implement it. One thing the OECD has done recently is launch a global framework for companies to report on their efforts to promote safe, secure, and trustworthy AI. This initiative monitors the application of the Hiroshima Process International Code of Conduct for Organisations Developing Advanced AI Systems. It provides a methodology for measuring and classifying different kinds of risks related to AI, which OECD governments have agreed to use. The aim here, she explained, is to build a shared risk management framework to allow the tracking of incidents linked to AI in a standardised way. This unique framework, however, will only work if industry embraces it and agrees to report on the AI risks and incidents it encounters.

"We want to take the incidents monitoring and methodology – which is important because it says 'here's what an incident is' and 'here are the categories of harm' – and standardise and globalise this information. [...] At the OECD, all 38 countries have agreed to it but we want it to be global.
If it is not global, it can't measure something useful."

Audrey Plonk,
Deputy Director of the Science, Technology,
Tech and Innovation Directorate, OECD

However, the OECD's role is different from that of AISIs, in that it is not a technical standards body: it helps policymakers understand standards and implement norms. In that sense, it bridges the gap between the technical standpoint and the political one. One thing the OECD could do, for instance, mentioned **Audrey Plonk** (OECD), is build on the vast amount of data they have gathered on what's happening in AI and translate it into action that can help the AISIs: *"I hope the safety institutes will come work with us, help us advance our risk management framework, and help inform us on what we can do to help them. Again, we are not a technical body, we are not going to write the testing standards, but we might help advise policymakers on how to implement it in their structures and in their regulations. This way, we can have both sides at the table: the technical and the policy side",* she added. The OECD, who has contributed to the international AI safety report led by Prof. Bengio, will also continue contributing to future versions of the report.

Conclusion

While global governance around AI safety is proving to be crucial, it also presents significant challenges. This roundtable highlighted the fact that a scientific consensus needs to be set in order to establish harmonised standards at a global level, and that this consensus around defining, evaluating and mitigating AI risks can only be reached through global cooperation.

During the AI Action Summit, multiple voices were raised that called for deregulation in the AI Sector in order to foster innovation. **Audrey Plonk** (OECD) disagrees with this narrative: *"I wish we could ensure that our future dialogue is not a binary one, meaning one in which safety and security isn't considered as a barrier to progress or as something that is going to inhibit or something that we can't have. The ideas that we can have both safety and prosperity should go hand in hand."* **Rebecca Finlay** (Partnership on AI) shares the same opinion. For her, it is thanks to concepts like responsibility and safety that innovation will be unlocked.

While the key focus of the discussion was on AI safety at the global level, the panelists underlined the need to think about it in an inclusive, globally diverse way. Indeed, marginalised countries may suffer in different ways than others from AI's potential risks. A challenge will be to take into account the voices from the Global South, which are often unheard.

AUTHOR
Astrid van de Blankevoort
Project assistant, Renaissance Numérique

Renaissance Numérique
Koburo, 35 rue Chanzy
75011 Paris
www.renaissancenumerique.org