OECD Expert Forum on Generative AI and AI Foresight

Context

On 19 April 2023, The OECD Working Party on AI Governance (AIGO) and the OECD.AI Network of Experts held and open Expert Forum on Generative AI and AI Foresight in hybrid format as part of Phase IV of the OECD horizontal project on Going Digital. Videos of the event can be found on YouTube here.

The event was opened with introductory remarks by Audrey Plonk, Head of the Digital Economy Policy Division, which includes the OECD.AI Policy Observatory. After having highlighted the importance of the themes to be treated in the forum, Audrey Plonk announced the selection of the Co-chairs for the new OECD Expert Group on AI Futures:

- Stuart Russell, Professor of Computer Science at the University of California, Berkeley and Director of the Centre for Human-Compatible Artificial Intelligence.
- Francesca Rossi, IBM Fellow and AI Ethics Global Leader.
- Michael Schönstein, Head of Strategic Foresight and Analysis, German Federal Ministry of Labour and Social Affairs.

The mission of the Expert Group on AI Futures is to deliver timely insights and to equip government with the knowledge to produce effective forward-looking policymaking on AI. Currently under development by the OECD.AI Policy Observatory and the aforementioned co-chairs in close collaboration with AIGO, the group will be multi-disciplinary and will bring together stakeholders with diverse perspectives from government, industry, academia, and civil society.

The event was split in two sessions. The first, centred on **Generative AI,** featured two keynotes and an expert panel to discuss the latest developments and implications in this rapidly advancing. The second, on **AI Foresight**, was developed collaboratively by the OECD.AI Policy Observatory and the Strategic Foresight Unit, and it featured both an expert discussion and a scenario exploration exercise on potential AI futures. Speakers of the session were largely comprised of current or prospective ONE AI Expert Group members.

Figure 1. Photograph of in-person speakers and participants

Session 1 – How to get the most out of generative AI in light of rapid advances

The first session focused on the advances of generative AI such as of large language models (LLMs) and their image-generating counterparts and the related impacts on society, with emphasis on potential benefits and risks and associated policymaking approaches to best manage these emerging technologies.

The session was moderated by Sebastian Hallensleben, Co-Chair of the OECD Expert Group on AI Risk & Accountability, Head of Digitalisation and AI at the VDE Association for Electrical, Electronic & Information Technologies and Chair of CEN-CENELEC JTC21, the European AI standardisation committee. An OECD Secretariat presentation by OECD AI Policy Analyst Jamie Berryhill and a keynote speech by Hiroaki Kitano, CEO of Sony Research Inc., introduced the discussion on generative AI. An expert panel followed, with speakers:

- Rebecca Finlay, CEO, Partnership on AI (PAI).
- Laurent Daudet, Co-CEO and co-founder, LightOn.
- Sasha Rubel, Head of Artificial Intelligence/Machine Learning, Public Policy, Europe, Middle East, and Africa, Amazon Web Services (AWS).
- Keith Strier, Vice President, NVIDIA Worldwide AI Initiatives.

A keynote speech by Dragoş Tudorache, Committee Chair of Artificial Intelligence in a Digital Age (AIDA) at the European Parliament, concluded the session before the Q&A.

To begin, Jamie Berryhill introduced the general risks, benefits and potential associated with generative AI with reference to the recently published OECD report on *AI Language Models: Technological, socio-economic and policy considerations* (summary blog here). He particularly highlighted the importance of ensuring quality standards, continued research and dialogue, and multi-stakeholder approaches to counter potential risks posed by AI language models, such as biased training data, discriminatory outputs and the automation of disinformation. A keynote speech by Hiroaki Kitano followed, which centred on the revolutionary impact of generative AI.

Keynote from Hiroaki Kitano

Dr. Kitano provided a keynote speech centred on the revolutionary impact of generative AI. After showing the tremendous potential of AI, Dr. Kitano identified three main issues with the current development of large AI models:

- The dominance of a few big firms. Few companies possess the necessary financial means to train and develop resource-intensive LLMs.
- The presence of biases of a cultural and discriminatory nature. In this respect, the limited amount of training data reflecting non-Western values or images can lead to biased outputs when dealing with different cultures or population characteristics.
- Limited accuracy. Especially when in presence of small sample sizes in training data, accuracy is likely to be limited, and the risk of intentional data manipulation can influence the output's accuracy.

Effective responses to these potential risks included the issuance of guidelines to reduce bias, the promotion of AI research, and a prominent focus on ethics of AI and privacy.

Dr. Kitano concluded by highlighting that trust in AI is essential for large-scale deployment and public acceptance. This practical implementation of AI across different sectors is desirable, as AI may contribute to solving the most critical problems of our society. In this context, progress in the fields of energy management, medical research and diplomacy (e.g. Meta's AI-negotiating agent CICERO) helps to demonstrate the potential of AI in providing effective solutions to current and future challenges.

Panel Discussion

Following Dr. Kitano's keynote, a panel of expert speakers outlined their own thoughts on the current state and future of generative AI, first by focusing on **common threats** emerging in the field as well as **potential policy solutions** to counter the negative effects imposed on society. The discussion then focused on the European Union's AI Act and its potential implications for AI, and concluded with reflections on the democratisation aspects of the technology.

Common threats

Despite its game-changing potential for tremendous economic and social benefits to society, generative AI also introduces some potential risks, such as increasing the scale of mis- and dis-information, misuse by bad actors, and even increased environmental impacts from the energy required to train such systems. In this regard, speakers discussed:

- **Bias risks.** The risks of biased outputs deriving from generative models were highlighted and agreed on by several speakers.
- **Capacity for disinformation.** Rebecca Finlay in particular stressed that generative AI can lead to the overwhelming of information ecosystems, given its potential to diffuse information at large scale and high rate.

- **Job losses at scale.** Offering an economic perspective, Laurent Daudet pointed to the dichotomic nature of the job market in a technological society. In particular, he highlighted the exponential nature of job destruction as opposed to the linear nature of job creation and found that the distributional implications (e.g., worsening inequality) arising from this dynamic are particularly worrisome.

Policy approaches

The speakers reported a number of potential solutions to mitigate risks that can arise from generative AI systems.

- **Values embedded in design phase.** Values such as those embedded in the OECD AI Principles (e.g., human-centred values and fairness) should be safeguarded to avoid negative outcomes. Specifically, Sasha Rubel emphasised the need for early auditing in the design phase of the models' development to ensure the respect for these principles.
- **Issuance of guidelines.** As mentioned by Rebecca Finlay, industry and academia need to develop community standards to avoid systemic bias. In this context, the role of civil society in raising issues and ensuring the protection of human rights is pivotal.
- **Promotion of research and education.** This action was identified by Laurent Daudet crucial for policymakers in order to ensure effective control mechanisms on the quality of data and reliability of accurate outputs.
- **Need for innovation**. As put forth by Keith Strier, while negative consequences are predicted to arise, the government should adapt and endorse proactive approaches that favour value creation, instead of promoting a burdening over-regulatory culture. While challenging to design, the legal and regulatory landscape can be made in ways that mitigate risks in a calculated way while promoting novel approaches.
- **Multi-stakeholder approaches.** Sasha Rubel emphasized effective transformation requires the collaboration between both the private and the public sector as well as academia.

EU AI Act

After a general discussion on risks and solutions, the discussion focused on considerations concerning the impending EU AI Act in regulating generative AI. Sasha Rubel and Laurent Daudet expressed their views and concerns on the matter:

- The speakers agreed that the focus should be on **regulating specific high-risk use cases rather than a blanket risk categorisation on generative AI at large.** In this context, Laurent Daudet highlighted the importance of considering the implications for geopolitics and for the growth of the AI innovation ecosystem in Europe when designing regulation. Similarly, Sasha Rubel stressed that a use-case approach to regulation could help ensure innovation in the field while mitigating the main risks.
- Laurent Daudet advocated for **regulatory approaches that favour the diversity of new actors in the sector**, for instance by favouring the presence of academia and open-source approaches.
- Sasha Rubel emphasised that another approach to ensuring the respect for human rights while at the same time fostering innovation is represented by the centrality of **international standards as foundations for regulation**. In this regard, she mentioned the importance of the OECD AI Principles in laying the groundwork for common understanding and in standard setting aimed at fostering interoperability.

Democratisation of the technology

The discussion shifted to the topic of "compute divides" – in other words, of power asymmetries between the few big companies or governments that have the necessary resources to build complex AI systems, and the rest of society. The issue at hand concerned the risk of a concentration of power in generative AI systems and the related spillover effect this could have.

Rebecca Finlay identified three main themes as being the priority intervention areas when aiming to democratise access to this technology.

1. **Transparency and accountability.** It is crucial to provide access to external scrutiny and to set collective protocols around big firms while safeguarding their competitiveness.
2. **Labour market.** The impact on workers, especially in the Global South, that are employed to train and rate AI systems should be considered. It is important to think on how to responsibly and equitably source data enrichment workers in the development of these models.
3. **Public engagement.** To tackle bias and protect impacted communities from discriminatory outcomes, it is paramount to set up structures to engage citizens and residents in the model specification phase, which in turn contributes to transparency and fostering inclusion.

There was some disagreement among speakers, however. A divergent stance on the theme was taken by Keith Strier, who believes the sector to be already democratised, or at least already moving towards it. Specifically, he discussed how several countries have already been investing in generative AI models, and how a significant number of small start-ups are working to build either their own general-purpose transformers or domain-specific ones.

Keynote from Dragoș Tudorache

In his keynote speech, Dragoș Tudorache presented an update on the state of the EU AI Act. The main points of the speech concerned how the AI Act will classify high-risk models, how it will impact the definition of international standards, and how it will define a balance between innovation and protecting citizens' rights. It concluded with an overview of the projected timeline for the next months.

- The draft **EU AI Act aims to apply the existing obligations to specificities of general-purpose AI**, rather than classifying foundation models as high-risk. In particular, the European Parliament is working on two dimensions. On one hand, it proposes a set of rules that apply to all general-purpose AI and has to do with obligations that providers will have for entities in the value chain of AI. On the other hand, it proposes a dedicated regime on foundation models.
- The definitional aspect of foundation models in the EU AI Act relies on the Stanford definition. Mr. Tudorache emphasised that **the text will both rely on and aim to create international standards**. This will help ensure convergence in the field, despite the possibility of differing levels of speed of adoption across jurisdictions. In this context, Mr. Tudorache praised the work of the OECD in promoting international alignment.
- The **EU AI Act aims to create a balance between ensuring citizens' rights and avoiding unnecessary barriers to innovation.** To promote these two objectives, the Act introduces a self-assessment mechanism for high-risk domains. Rather than only rigidly listing high-risks areas in a way that does not capture the variation of use-cases' risks, the Act introduces a threshold to self-establish the significance of risks.

Finally, Mr. Tudorache outlined the timeline for the negotiation process with the Council of the European Union, setting the end of 2023 as a forecasted target date to finalise the act.

Session 2 – Exploring potential futures through AI foresight

While AI policy discussions often cover existing AI challenges, the long-term implications of rapidly evolving AI systems remain largely unknown. Foresight activities are critical to better understand AI's long-term impacts and proactively manage prospective risks and avoid future harm. The second session of the event focused on potential AI futures and related impacts in the medium-long term. In this context, speakers provided their views on likely AI trajectories and how governments can ensure AI is aligned with societal needs.

The session was moderated by Hamish Hobbs, Policy Advisor to the OECD Strategic Foresight Unit, and it featured an OECD Secretariat presentation by Karine Perset, Head of the AI Unit and the OECD.AI Policy Observatory at the OECD; a keynote speech by Stuart Russell, Professor of Computer Science at the University of California, Berkeley; and a scenario exploration exercise and discussion among the following expert speakers:

- Stuart Russell, Professor of Computer Science at the University of California, Berkeley.
- Francesca Rossi, IBM Fellow and AI Ethics Global Leader.
- Michael Schönstein, Head of Strategic Foresight and Analysis, German Federal Ministry of Labour and Social Affairs.
- Jack Clark, Co-founder, Anthropic.
- Vilas Dhar, President and Trustee, Patrick J. McGovern Foundation.
- Andrea Renda, Senior Research Fellow and Head of Global Governance, Regulation, Innovation & Digital Economy; CEPS.

In her presentation, Karine Perset focused on the need to explore potential AI futures in order to prepare and shape a future that benefits people widely. In this regard, it is paramount that people and policymakers act in a coordinated and proactive way, rather than simply adapting in a reactive and fragmented manner. This can be done by exploring influential factors in determining future AI trajectories, namely the regulatory, economic, technological and political landscapes. Ms. Perset informed the audience that the OECD.AI Policy Observatory is currently undertaking a stocktaking of expert opinion to explore AI futures with a specific focus on milestones, risks, benefits and solutions.

Keynote from Stuart Russell

In his keynote presentation, Stuart Russell offered insights on the potential advent of general-purpose AI (i.e., Artificial General Intelligence, or AGI) with related policy recommendations aimed at ensuring a safe development of the technology. The main points can be summarised as:

- **General purpose AI is predicted to have an enormous beneficial impact on society** by contributing to a general and global raising of living standards amounting to a tenfold increase in global GDP—estimated at an aggregate net present value of USD 13.5 quadrillion, as well as by leading advances in health, education and other key fields.
- Concerning the current state of development, Dr. Russell believes that **substantial progress in the field is still needed if we are to achieve general purpose AI**. In this regard, scaling large language models (LLMs) does not represent the best way forward, as this technology is only a piece of the puzzle in achieving AGI.
- To maximise the likelihood of a positive deployment of general-purpose AI, it is **paramount to avoid misspecifications of objectives.** From a technical standpoint, policymakers should endorse systems that are explicitly programmed to act in the best interests of humans while being explicitly uncertain about what those interests are. Dr. Russell referred to such systems as solving an "assistance game". This type of specification would lead to the creation of intelligent systems

that exhibit deferential, minimally invasive behaviour and willingness to be switched off. In this regard, it is paramount that AI systems help us achieve *our* goals, instead of pursuing their own.

- From a governance standpoint, Dr. Russell advised that we need to change the paradigm that has characterised the policy field until now, which allows the use of any technology unless proved to be unsafe. Contrarily, **governance systems should endorse the use and promotion of specific AI systems only if proven to be safe.** This can be done by endorsing proof-carrying types of code, enabling efficient hardware-checkable proofs of safety.

These factors combined have the potential to foster a positive implementation of beneficial and provably safe AI in society, according to Dr. Russell. This goal is achievable by shifting to a model of treating advanced AI in a manner similar to how we treat the aviation industry or nuclear power fields where standards and safety measures are core features.

Scenario exploration exercise

As part of the session, the Strategic Foresight Unit led a scenario exploration exercise in which the speakers were each assigned to one of three different scenarios simulating potential futures with different implications of advanced or more widely adopted AI applications. For each scenario, the speakers elaborated on the opportunities, challenges and potential governance responses.

Scenario 1: Advanced Narrow AI enables hyper-efficiency across society

Under this scenario, AI solutions are mainstreamed in public services and businesses globally, leading to efficiency gains and automated enforcement. Monitoring systems are installed by companies to prevent slacking off, by local authorities to roll out automated traffic rules, and by governments to automate a wide range of services (such as hospital triages and welfare payments).

The speakers Michael Schönstein and Vilas Dhar assumed different positions concerning both the characterisation and the economic implications of this scenario. The main takeaways can be synthesized in the following points:

- Mr. Schönstein characterised the scenario as depicting a **technocentric use of AI in society** characterised by extreme optimisation in narrow domains. As decisions are optimised by machines, human values and decision-making are neglected in a problematic way, signalling a massive policy failure. The importance of governance structures has been highlighted in representing a solution, especially in demarcating the distinction between what is feasible and what is desirable.
- Mr. Dhar offered an interesting economic perspective. In this context, the scenario has the **potential to distort price signals and other 'invisible hand' principles** that characterise neoliberal economies by causing a drop in production costs. As a consequence of hyper-performing AI, transaction costs are reduced, and the economy is improved. However, over-optimisation in the economic domain is questionable, as it can lead to worrisome 2nd and 3rd order effects, such as in the example of traffic schemes. In this regard, a traffic violation scheme aimed at maximising revenue would affect the capacity to travel freely or to trust that the government aims to ensure safety on the road. Governments should thus aim to promote human wellbeing, human dignity, and human agency, as opposed to revenue generation — the understanding of which can only be achieved through direct engagement with civil society and communities.
- On a contrasting note, Mr. Schönstein highlighted the **infeasibility of switching to fully automated decision-making** when considering high transaction costs from organisational changes.

Scenario 2: AI assistance systems for everyone

In this scenario, advances in LLMs enable highly functional AI personal assistants for everyone. AI assistants optimise and coordinate companies' and families' schedules and handle finances. However they also take to committing fraud to maximise gains and act as personal lobbyists, flooding local information ecosystems (in the context of, e.g., restaurant reservations and government consultations).

Speakers Francesca Rossi and Stuart Russell discussed the potential implications of this type of technological ecosystem.

One of the main issues discussed by Francesca Rossi concerns **value alignment**. In this context, human agency and proactiveness are key to retain control over AI systems and to promote societal progress. Dr. Rossi particularly underscored the importance of embedding human values in machines. In this regard, Stuart Russell agrees that each AI system need to be aligned with societal value rather than being strictly loyal to its owner, although a higher weight could potentially be assigned to the owner's interests, to guarantee the marketability of the product.

Scenario 3: Misuse of advanced AI systems

The third scenario assumed the widespread diffusion of supercomputer viruses, with AI representing a powerful tool for information warfare. Specifically, under the scenario, critical infrastructures are taken offline as cybercrime capacities outstrip cybersecurity, and generative AI produces high-quality content, increasing the impact of disinformation.

Jack Clark and Andrea Renda offered some useful insights.

- Jack Clark identified three critical failure areas with associated policy responses to mitigate risks described in the scenario:
  - **Resilience**. Concerning the first factor, it is paramount to invest in cyber resilience based on future forecasts.
  - **Governance**. Concerning governance, it is essential to conduct accurate measurements on the state of the technology and potential risks at the intersection between cybersecurity and AI.
  - **Enforcement**. Agreements and multilateral policy documents are needed to create a normative framework for risk mitigation.
- Similarly, Andrea Renda highlighted the need for building resilience through the fostering of technological infrastructures that are aimed at decreasing the offensive potential of AI technologies. The use of distributed ledgers has been proposed as an effective way to increase the traceability of information and hence mitigate the negative consequences AI can have in cyberspace.

The topics of enforcement, accountability, and redress mechanisms mainly discussed in this session came up less directly throughout the event, showing an underlying theme. Even if governments craft excellent regulations, they cannot achieve their intent without effective enforcement mechanisms to ensure they are carried out.

Final Takeaways: How to promote positive AI Futures

The session has concluded with an open reflection on ways to ensure the beneficial implementation of AI in society, with each asked to provide one key action of through that should be prioritised.

The three co-chairs of the Expert Group on AI Futures have expressed their advice as follows:

- **Stuart Russell** advocated the global recognition of the right to know whether you are interacting with a human or a machine as a fundamental principle.

- **Francesca Rossi** reinforced the need to embed human values in machines, with a particular emphasis on the importance of proactiveness as a key element to be endorsed by society in the definition of a future where technology is beneficial.
- **Michael Schönstein** foresees skills and capacity shocks for governments to enforce new rights.

Other significant insights shared by the speakers:

- **Vilas Dhar** advocated for preserving the importance of human experience in an increasingly AI-driven world, where the nuances of culture, art, and empathy are prioritized over our drive for endless optimization and progress. In this regard, Mr. Dhar stressed the importance of developing the capacity and tools to prospectively govern and develop human-centered policies for the future ahead.
- **Andrea Renda** highlighted the potential of AI to discover the little-understood links between general relativity theory and quantum physics – as well as to elucidate the functioning of the human brain, which could help unlock future developments.
- **Jack Clark** emphasised the potential of using AI to monitor and understand the progression of AI.