OECD   OECD.AI Policy Observatory

AI Foresight: Anticipating what lies ahead

The context for Artificial Intelligence (AI) in Work, Innovation, Productivity and Skills (WIPS)

The 2023 International Conference on AI in WIPS was held on 27-30 March 2023. It brought together leading voices from public policy, academic, industry, technical, and civil society communities to discuss how AI affects key societal and economic areas, and how policymakers can respond. The event was opened by OECD Deputy Secretary-General Ulrik Knudsen and Nermin Fazlic, State Secretary of the German Federal Ministry of Labour and Social Affairs. The OECD.AI Policy Observatory organised two sessions within the conference. The first took place on the first day and focused on how countries can come together to manage AI risks. The second, the focus of this summary, took place on the third day and discussed how future AI systems may progress in the medium and long term - including the possible advent of Artificial General Intelligence (AGI)—also referred to as General Purpose Artificial Intelligence—and the potential implications to economies and societies.

Anticipating what lies ahead

In his keynote speech to help open the event, Stuart Russell, Professor of Computer Science at the University of California, Berkeley, acknowledged the vast potential of AI for good but also cautioned against the deployment of machines at a global scale that may or may not have their own internal goals, whose internal operations we do not understand, and whose capabilities we cannot predict.

Professor Russell's remarks can be seen as an effective synthesis of the main themes discussed in the session "AI Foresight: Anticipating what lies ahead". The session took the form of a panel discussion and was moderated by Marjory Blumenthal, Senior Policy Researcher at the RAND Corporation and Principal at MSBlumenthal, LLC, with speakers:

- Pam Dixon, Founder and Executive Director of the World Privacy Forum;
- Hamish Hobbs, Policy Advisor at Longview Philanthropy and to the OECD Strategic Foresight Unit;
- Nicklas Lundblad, Head of Global Policy and Public Affairs at DeepMind;
- Dunja Mladenić, Senior Researcher and Project Leader at the Jožef Stefan Institute; and
- Michael Schönstein, Head of Strategic Foresight and Analysis at the German Federal Ministry of Labour and Social Affairs.

The session centred on four main themes: **key milestones** in AI development in the medium to long-term; potential **risks** posed by AI in the next 10 years; the **role to be played by the OECD** in the context of AI foresight; and potential **trade-offs** between values and performance of AI systems.

Throughout the session, speakers agreed on the cruciality of quality data, the development of algorithms and methods that can take context into account, the importance of the scale of processing and computing power, the need for explainability and interpretability, and the role of generative AI models as a translational interface to create new types of access that may impact future developments. Speakers also reflected on how policymakers can use strategic foresight to anticipate and act to shape potential AI futures and keep up with technological developments.

Safety and security and risk mitigation was a recuring topic, as well as the need for measuring complexity, the future of domain-specific large language models (LLMs), and the importance of involving all stakeholders and actors in the discussions on the future of AI.

The session and its themes are well aligned with the objectives of the new OECD Expert Group on AI Futures, which was kicked off with the announcement of co-chairs Francesca Rossi, Stuart Russell, and Michael Schönstein shortly after the AI WIPS event.

Key tech advancements and milestones

Speakers identified key advancements characterising the near to medium-term future as being **federated learning** (a machine learning technique that relies on the use of multiple decentralised datasets), **quantum computing** (a technology using principles of quantum mechanics to solve complex problems), **DNA computing** (a branch of biomolecular computing which uses DNA to complete arithmetic operations), **specialised AI chips** (computer chips designed to perform and support machine learning techniques), embedded AI (the application of machine learning in software), and **user interfaces** (design interfaces for human-AI interaction). However, the need for **quality data** and robust **safety** measures will continue to be important. In particular:

- Michael Schönstein pointed to the importance of applying federated learning in contexts characterised by small datasets. Other technological advancements in the fields of quantum computing, DNA computing, and specialised chips were highlighted by Dunja Mladenić as representing fundamental progress in managing complexity and performance and effectively dealing with diverse data. Furthermore, a critical factor of the future that was highlighted is the application of existing models to different specific domains.
- A noteworthy common topic emerged throughout the discussion referred to data quality. According to the speakers, the future in the field will be characterised by the continuous identification, generation, and/or research of quality data to be used in complex ecosystems. In this regard, generative models would play a growing role in increasing access to different types of both data and data structures.
- Relatedly, advances in the governance of data ecosystems and in public data infrastructures were other two factors deemed as fundamental milestones in ensuring progress at a systemic level. In particular, Nicklas Lundblad stressed the need for developing institutional frameworks that are aimed at promoting system advancements and at fostering interoperability.
- A perspective shared by Pam Dixon was the crucial importance of safety structures (such as adversarial testing) in guaranteeing trustworthiness. As she mentioned, the rate of progress in the safety field in the next 10 years will be a crucial determinant of ensuring reliable AI systems in the long term.

Risks posed by AI in the short-term and potential mitigation options

Among the top risks posed by AI in the next 10 years, the speakers agreed on some common threats, namely a **lack of algorithmic and systemic transparency**, the **societal misconception** of AI capabilities, the **alignment problem**, an AI **development race** among corporations and governments, **exacerbating inequality**, and **network risks** (i.e. risks deriving from the unknown effects of increased interdependencies of AI systems and potential cascade failures). They also highlighted the important role to be played by **digital identities**. In particular:

- As discussed by Dunja Mladenić, lack of transparency in the field can be the consequence of closed systems where a small number of players have access to the technology and its operations, models, and data. To effectively respond to this threat it would be necessary to both work towards an idea of open science and to effectively increase transparency, especially concerning instances in which people are subject to algorithmic decisions.
- Misconception of AI systems can arise as humans anthropomorphise AI – attributing to AI biological traits such as intentions and comprehension. Nicklas Lundblad highlighted how “evolutionary traps”

might play a role in creating common misconceptions. As from an evolutionary perspective we tend to associate competence with comprehension, and we can easily be misled by attributing biological characteristics to non-biological AI systems. To tackle this issue, it is essential to promote outreach activities and basic understanding of what AI is and is not.

- The alignment problem refers to AI actions that do not align with intended goals and preferences. In particular, Hamish Hobbs and Nicklas Lundblad underscored the need to create new frameworks able to capture the increased autonomy of future AI systems and in turn effectively mitigate the deployment risks that may arise.
- A further critical risk raised by Pam Dixon is that of an AI race in the context of LLMs. A race in the development of the best-performing algorithm can, for instance, lead to worrying exploitation of patients' data in the healthcare sector. Moreover, with continued advancement in LLMs, flaws will decrease but so will the ability to identify them. The solution to the risks posed by a race would be to slowing down progress and putting in place mitigation strategies and policies commensurate with the potential risks and impact of development and deployment.
- Michael Schönstein discussed the risk of AI exacerbating economic inequality. As AI applications may have higher returns when used by higher skilled workers, they are liable to reinforce existing divisions in the financial domain in the near term. Mitigation can be achieved by focusing on upskilling of workers and the general population. In this context, speakers also underlined how AI systems impact the economy by creating demand for deep domain knowledge, as it is necessary to judge the usefulness and accuracy of AI-generated outputs.
- A common conception outlined during the session, is that assuming an ecosystem approach and relying on red teaming are key ways to mitigate network risks. Pam Dixon in particular emphasised the need for extensive adversarial testing for high-impact systems.
- Another key concept in this context that was mentioned by several speakers was the idea of a trusted digital identity to tackle disinformation. Pam Dixon highlighted how digital identities can lead to challenging and potentially harmful consequences for the population if abused. Hence, the need to embrace an ecosystem approach where the impacts of the technology on human rights and privacy are empirically assessed with the creation of benchmarks and milestones.

The role of the OECD in promoting positive AI futures

The speakers agreed on the fundamental role that the OECD both can and should tame in promoting beneficial practices in AI foresight. It can do so by:

- **Providing trend analyses** in AI adoption based on empirical evidence, for instance through an AI Incident Monitor or by promoting workshops on generative AI. Such items can both raise awareness of current risks and opportunities, as well as help detect weak signals to help anticipate future events.
- **Drafting and promoting guidelines on responsible AI**, while providing evidence-based advice for governments concerning best-practices and limitations in the field. In this context, Hamish Hobbs highlighted the prominent role to be played by the OECD Expert Group on AI Futures. Guidelines can for instance be applied to the field of randomised control trials (RCTs) in empirical AI assessments, as suggested by Michael Schönstein, and by evaluating the transferability and comparability of results.
- **Promoting diversity**. Another common insight that emerged throughout the session, and particularly emphasised by Pam Dixon and Nicklas Lundblad, concerned the problem of the prominence of 'Global North'-centred teams of experts. Hence, inclusiveness should be promoted, by endorsing multi-stakeholder approaches and by promoting expert groups' diversity on multiple

dimensions. In this regard, the OECD is seen as a valuable player, capable of promoting heterogeneity in the field.

- **Supporting beneficial cross-border data flows**. In particular, the role of guidelines has been identified as paramount in promoting beneficial cross-border data flows and regulations. This can help foster quality and trustworthy data capabilities, as well as increasing the applicability of foresight activities to policymaking.

Finally, discussion members highlighted how the neutral role that characterises the OECD can be effective in bringing actors together to collaborate on key issues under evidence-based frameworks. Dunja Mladenić, in particular, highlighted the agility of the OECD in supporting AI policy work as its focus on legal and technology policy combined enables it to effectively provide recommendations to governments.

Trade-offs between explainability and performance of AI systems

In the context of AI performance trade-offs, speakers agreed that quality and efficiency do not necessarily need be sacrificed to guarantee values such as explainability. Explainability of algorithmic systems can be achieved in a flexible manner that does not hinder accuracy. Tools such as adversarial testing can help in achieving accuracy, security, and efficiency in ways that still ensure positive economic, societal, and environmental impacts.

Conclusion & key takeaways

Overall, the session has provided important insights on the progress, risks, mitigation strategies and role of actors in the context of AI foresight. The speakers discussed their thoughts on the most important focus areas going forward, which can be synthesized as:

- The role of adversarial testing and institutional frameworks in ensuring reliable, safe and trustworthy systems.
- Ensuring systems' alignment and avoiding perilous races to progress, represent the policy priorities for the future advancement of AI.
- As the technology evolves into more complex, multi-faceted and interoperable systems, it is necessary that we promote novel guidelines and frameworks to effectively deal with them, using evidence and multi-stakeholder approaches as foundational principles. The role of the OECD in this regard is to help pave the way to informed and effective policymaking advice in the AI foresight field.
- AI transcends national borders. No single country, organisation or stakeholder can solve problems and challenge alone. Collaboration will also be needed to secure the full potential benefits of AI. Intergovernmental organisations have a key role, and the OECD is well positioned in helping policy makers develop common visions and solutions.

The session concluded with a quote from Edward O. Wilson shared by speaker Nicklas Lundblad (DeepMind), "*The real problem of humanity is the following: We have Paleolithic emotions, medieval institutions, and godlike technology.*" The quote is exemplary of the takeaways of the session. As technology advances and increases its potential to impact society, it is paramount to develop guidelines, adapt institutions and create new frameworks able to mitigate the risks and foster beneficial practices.