OECD

OECD.AI Policy Observatory

OECD Expert Group on AI Futures (14 November 2024)

Background

Immediately following the GPAI 2024 Fall Plenary, the OECD Division on AI and Emerging Digital Technologies and the OECD Strategic Foresight Unit (SFU) convened for the sixth meeting of the Expert Group on AI Futures on 14 November 2024. The expert group is a core component of the OECD workstream on AI Futures.

The Expert Group is led by three co-chairs:

• 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 General Digital Policy - Federal Chancellery of Germany.

The full composition of the Expert Group is available here.

Introduction and context

This meeting was held in hybrid format immediately following two days of meetings of GPAI Plenary. While the meeting discussions were held mainly among members of the Expert Group on AI Futures, GPAI delegates and other members of the broader OECD-GPAI Expert Community were welcome to attend and participate. The meeting was held under the Chatham House Rule.

Karine Perset, Acting head of the OECD Division on AI and Emerging Digital Technologies, opened the meeting by welcoming attendees and giving brief introductory remarks. She also provided an organisational update on the new integrated partnership, as the OECD and GPAI have joint forces.

Following Karine's introduction, Michael Schönstein took the floor to guide attendees through the meeting's agenda, which focused on:

1. A recap on the Expert Group's activities and plans.
2. Presentation and discussion on the International Scientific Report on the Safety of Advanced AI.
3. Presentation and discussion on the survey on thresholds for advanced AI systems.
4. Presentation and discussion on high-level project proposals for 2025.

Attendees were encouraged to share additional thoughts via email at ai@oecd.org if unable to contribute fully during the meeting or if they had new thoughts afterwards. This summary comprises both the meeting as well as any subsequent input received. The co-chairs facilitated discussions on the agenda items, as detailed below.

A recap on the Expert Group's activities and plans.

Michael Schönstein remarked that in its first year, the Expert Group identified, analysed and prioritised over 100 potential future AI benefits, risks, and policy imperatives. He also announced that the paper summarising the conclusions of their work would be published, and the scenarios developed and discussed by the group were also used as a key input to the OECD's Global Strategy Group meeting which discussed the Futures of Global AI Governance. Looking ahead, he noted that the group's focus will shift to deep dives on specific priorities identified by the group and GPAI, with plans to create space for more direct expert engagement by forming sub-groups to lead the work on specific deep dives.

Presentation on International Scientific Report on the Safety of Advanced AI

Presentation on International Scientific Report on the Safety of Advanced AI

Prof. Yoshua Bengio, Founder and Scientific Director of Mila and Professor of Computer Science at the University of Montreal, along with Benjamin Prud'homme, Vice-President of Policy, Safety, and Global Affairs at Mila, presented the draft International Scientific Report on the Safety of Advanced AI. The presentation included:

- **Context and Mandate of the Report**: The report is based on the 2023 Bletchley Declaration, which emphasises the need for an inclusive, international, and independent report on the capabilities and risks of advanced AI. The objective of the report is to support policymakers with expert information and evidence-base. The Interim Report, focusing on General-Purpose AI, was presented at the AI Seoul Summit in May 2024.
- **Current Process:** Feedback has recently been collected from civil society and industry, and the final report is to be delivered as part of the AI Action Summit in February 2025.
- **Findings at a quick glance:** General-purpose AI offers great opportunity, if properly governed. But there are risks such as malicious use risk, risk from malfunctions, and systemic risk. Technical methods exist to reduce risks, but all have limitations. Experts disagree over some of these risks, mainly due to differing expectations about how rapidly AI capabilities will advance. Better understanding of general-purpose AI is necessary.

Discussion among meeting participants

Overall, participants expressed appreciation for and interest in the report. Experts raised the question of when open-source approaches are appropriate given the potential for trade-offs between safety and openness. It was noted that the risks and benefits of openness vary based on context and need to be assessed on a case-by-case basis. Systematic approaches to evaluating these potential trade-offs were highlighted as an important area for further research.

Experts also highlighted that approaches are needed to integrate the findings from scientific assessments like this International Scientific Report on the Safety of Advanced AI into policy processes, to empower policymakers with the best available information. Experts discussed the need for different types of reports, with a clear separation between those focussed on the science of advanced AI and those focussed on policy responses based on this science.

Presentation on the survey on thresholds for advanced AI systems

Presentation on the survey on thresholds for advanced AI systems

Prof. Robert Trager, Director of the Oxford Martin AI Governance Initiative, Jonas Schuett, Senior Research Fellow at the Centre for the Governance of AI, and the Secretariat (Ms. Eunseo Dana Choi) presented preliminary results from an OECD survey and public consultation on thresholds for advanced AI systems.

Presenters noted that the AI Seoul Summit 2024 highlighted the critical role of thresholds in AI governance. Thresholds are included in major AI regulations and standards such as EU AI Act. The expert survey and public consultation sought to assess expert and public views of the role and value of thresholds as a policy tool to manage AI risks. The survey had 166 participants from the OECD Network of Experts on AI, and 45 participants contributed to the public consultation.

The findings from the survey and consultation revealed several areas of agreement among participants. There was consensus on the importance of multi-stakeholder involvement, emphasising that thresholds should not be determined solely by AI companies or solely by governments, but rather through broader engagement with the private sector, academia and civil society. Participants also agreed on the need for multiple types of thresholds, noting that different thresholds should serve different purposes based on different metrics. Additionally, several desirable criteria for thresholds were identified, including that thresholds should be externally verifiable, enforceable, future-proof, and justified. Challenges in setting and evaluating thresholds were also highlighted. These included conflicts of interest among stakeholders involved in threshold setting, measurement challenges due to the lack of reliable methods for estimating risks and assessing model capabilities, and disagreements on the appropriate actions to take when thresholds are breached.

Discussion among meeting participants

Participants suggested that further consideration is needed regarding when and if thresholds are an appropriate tool for governance in any given situation, highlighting that any use of thresholds should have a well-reasoned basis. For instance, a threshold might be appropriate where risks rise substantially above a certain threshold value, as exemplified by the use of thresholds in speed limits for vehicles, but may not be appropriate in some other contexts. Additionally, participants highlighted that that thresholds will need to be contextual, considering the implications of specific applications in specific sectors. Some experts noted the importance of distinguishing between risk tolerance thresholds and the methodology for assessing the level of risk, with risk tolerance thresholds being a normative consideration while risk levels can be subject to empirical assessment. Some experts also pointed to the significant gap between what we know how to measure and what we would like to measure, highlighting need for further research on relevant metrics.

Discussion on High-level project proposals for 2025

Presentation on High-level project proposals for 2025

Yuko Harayama, the Head of Tokyo GPAI Expert Support Centre, presented high-level project proposals for 2025. Proposals included:

- **Agentic AI:** Analyse the implications, including benefits and risks, of agentic AI systems and develop a framework for classifying them based on their characteristics.
- **AI Red Lines in High-Risk Industries:** Taking stock of red lines in high-risk sectors and identifying both technically achievable and politically feasible AI red lines in different contexts.
- **Safe AI by Design:** Exploring various approaches to AI safety by design, including creating a knowledge repository of tools and good practices.

Yuko concluded by inviting feedback on the proposals. These proposals are being reviewed in parallel by GPAI Members.

Discussion among meeting participants

Participants expressed appreciation for the high-level project proposals for 2025. There was broad agreement that the Expert Group should focus upon its comparative advantage, which includes the application of futures methods such as scenarios analysis. With this in mind, the discussion coalesced around two key areas of future work for the group:

1. **Scenarios exploring how AI capabilities may evolve in the future.** These scenarios would explore potential future technical developments in AI and how these developments could translate into task-specific or general capability improvements for future AI systems. These scenarios would attempt to rigorously define different potential future capabilities of AI systems, and synthesise the best available evidence to explore if, when, and how these capabilities might emerge. These capabilities scenarios could then be used to inform policymaking today, to ensure governments are prepared for plausible AI capabilities scenarios.
2. **Scenarios exploring potential future sectoral/topical impacts of AI.** These scenarios would explore the potential impacts of future developments in AI on specific sectors and stakeholder groups, such as healthcare, impacts on children, education, media, interpersonal communication, or energy. These sectoral or topical scenarios could build on the capabilities scenarios, exploring how different AI capabilities could impact specific sectors or topics. These scenarios could then be used to guide specific sectoral governance approaches or to draw broader insights that may be applicable across sectors.

Key elements to capture in proposed scenario exercises

Capturing people’s perception

Participants emphasised the importance of capturing people's perceptions and expectations about the future, noting that AI Futures projects under the UN revealed gaps in understanding this dimension.

Capturing human and societal change in AI futures

One expert highlighted the need to explore human and societal changes that might emerge in AI futures, including potentially unusual changes in society as people adopt current or foreseeable technologies. One potential example was whether future interactions between humans might be frequently mediated by bots, in the same way they are currently frequently mediated by other digital tools.

Geographical differences

Participants suggested considering geographical differences, including differences between developed and developing contexts. For instance, sectoral deep dives could consider the future impact of AI on a given sector in developing countries.

Using existing OECD tools

Participants noted that the OECD.AI Policy Observatory already tracks many aspects of the AI ecosystem, and these metrics could be used to inform scenario development.

Explore all types of future developments

Experts highlighted that the scenarios should not focus exclusively on risks or benefits, but cover the full range of potential future developments and implications of AI systems.

Importance of understanding technology first

One expert stressed the need to first understand technological developments and their capabilities, as this foundational knowledge is critical for predicting potential impacts and identifying relevant policy issues. Experts highlighted potential technological developments to explore that included capability gains from continued scaling of models, specialised datasets, neurosymbolic AI, AI agents, and the development of smaller, more efficient AI models.

Additional perspectives on topics in project proposals

Agentic AI

Participants agreed that agentic AI is a relevant topic. They suggested that any analysis of the topic would need to be clearly defined and scoped. For instance, they noted that the Expert Group could explore the implications of plugins and integrations for language models. These may or may not meet different definitions of an agent, but can nonetheless enable LLMs to interact with and influence the physical or virtual world. Given the unpredictable behaviour of LLMs, the impacts of these integrations could merit policy attention. Another participant suggested that rather than starting from scratch, the OECD's existing Framework for the Classification of AI systems could be updated to include consideration of agentic AI.

Red lines

One expert highlighted that the proposal about red lines is particularly timely for 2025, as the provisions in the European Union's AI act will come into force in February, and there will be an interpretive process to figure out how and where they apply. The expert also suggested exploring the externalities of imposing bans, noting that red-lines provisions could cause potentially useful applications to be restricted. The expert suggested exploring how to effectively prohibit harmful practices minimising unintended consequences.

Other suggestions on processes and working methods

Handling disagreements

Participants suggested that it would be beneficial for the group to have a clear understanding and processes for handling internal disagreements, including how to proceed in such situations and ensure publications can reflect expert disagreement. In response, another expert noted that one approach to resolving such issues involves gathering experts' opinions and presenting a distribution of responses rather than relying on consensus.

Avoiding duplication

Participants emphasised the importance of considering the unique contributions the group can make while avoiding duplication, especially given the numerous projects currently underway within GPAI.

Utilisation of the OECD.AI Policy Observatory

Participants pointed out that many proposed ideas could be implemented using the OECD.AI Policy Observatory, which tracks the entire AI ecosystem, including academic publications, AI incidents and hazards, venture capital investments, and policy developments. These resources could be leveraged for scenario development or for early indicators of emerging trends.

Conclusions

The discussions coalesced on the comparative advantage of the Expert Group in delivering scenarios analysis to envision AI futures. These scenarios could explore both potential future AI capabilities, and the sectoral impacts of these capabilities.