OECD Expert Group on AI Futures (17 March 2025)

Background

The OECD Division on AI and Emerging Digital Technologies and the OECD Strategic Foresight Unit (SFU) convened for the seventh meeting of the Expert Group on AI Futures on 17 March 2025. 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 virtual format. The meeting was held under the Chatham House Rule.

Luis Aranda, Senior Economist/Policy Analyst of the OECD Division on AI and Emerging Digital Technologies, opened the meeting by welcoming attendees and giving brief introductory remarks.

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

1. Plans for 2025 based on last meeting
2. Presentation and discussion on the AI capabilities deep dive
3. Presentation and discussion on the agentic AI deep dive

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.

Plans for 2025 based on last meeting

Michael Schönstein remarked that in previous meetings, the group had agreed to focus on exploring scenarios in two key areas:

- Deep dives into how AI may evolve in the future
- Potential future sectoral and topical impacts of AI

Michael noted that the group would first conduct deep dives on how AI may evolve in the future, which could then inform discussions on sectoral and topical impacts. During the meeting, the group planned to launch two deep dives and create space for more direct expert engagement by forming sub-groups to lead the work on specific deep dives:

- AI capabilities scenarios – exploring plausible AI capability scenarios
- Agentic AI – examining near-term futures of agentic AI systems

Presentation and discussion on the AI capabilities deep dive

Presentation on AI capabilities deep dive

Michael Schönstein presented a deep dive on AI capability scenarios, outlining the following proposals:

Aim

- Define a set of mutually exclusive AI capabilities scenarios, ranging from a capabilities plateau to transformative advances.
- Identify evidence for or against the plausibility of each capabilities scenario materialising.
- Assess the relative likelihood of each scenario and each scenario's policy implications.

Scenario structure

- Proposing to explore four capability scenarios: plateau, incremental advance, substantial advance, transformative advance.
- Each scenario could explore three dimensions of capabilities: task-specific performance, generality/range of tasks and continual learning.

Discussion on AI capabilities deep dive

Stuart Russel opened the floor, inviting feedback on the following discussion questions.

1. Are these the right levels for AI capability scenarios?
2. What is needed to make the capability scenarios rigorous and meaningful?

Participants provided feedback on the proposed framework and raised several important points:

Use a scenario-based approach
The scenario-based approach was generally well-received, with participants emphasizing the importance of using evidence-based and scientific analysis and agreeing that the proposed capability levels covered the range of plausible alternative futures. Participants highlighted the need to be systematic across the scenarios to enable meaningful comparisons.

Ensure appropriate scoping

Several experts noted that the proposed scope was quite limited, with scenarios following a largely linear progression of increasing AI capabilities rather than capturing a range of possible variables. Socio-technical, geopolitical, economic, and governance factors were identified as important variables. Experts noted that AI futures could look very different depending on variables such as the levels of value alignment, reliability, or agency of AI systems, the degree of centralisation or diffusion of leading AI models, levels of AI adoption, and the types of AI governance models that are implemented. Others noted that it would be difficult to infer policy implications and governance guidance without considering the broader context in which future AI systems may be developed or deployed.

Some experts highlighted that it will be important to very carefully and deliberately constrain which variables the scenario analysis will consider. These experts suggested that scenarios about AI need to be bounded to be rigorous and comparable.

To address both of these priorities, a staged approach was proposed. This approach would involve initially developing a set of more limited technical capability scenarios. For practical purposes, the extent and context of the deployment and access to these systems could initially be excluded from the scope. Future work of the expert group could then consider additional relevant variables, including the contexts in which these capabilities may be developed and deployed.

Focus on frontier AI systems

To further focus the scope of the scenarios exercise, it was suggested that the exercise's scope be limited to the capabilities of frontier models as they may exist at future time points. Experts suggested that identifying the potential frontier of capabilities may be easier to predict and assess compared to levels of AI deployment and the societal impacts of AI systems, which involve additional uncertainties.

Assess policy and governance implications

The importance of addressing policy and governance issues across different scenarios was highlighted.

Be precise

Experts highlighted the need to be precise about the variables explored in the scenarios, such as being clear in definitions of performance, tasks, and capabilities. Experts noted that frontier AI systems at any of the four proposed scenario levels could be composed of many different stacks of capabilities, and the impacts of these systems would depend strongly upon which exact capabilities were present. For instance, one expert questioned whether robotics performance should be part of the discussion of task-specific performance. Another expert noted that speed of task completion could be a relevant performance metric to consider. A structured approach to capability selection was recommended, focusing on identifying distinct domains (e.g., education, construction, media) to facilitate a more targeted discussion.

Adjust timelines

The choice of 2035 as the target year was questioned. Experts noted that AI advances could be extremely rapid, and policymakers would benefit from information about potential advances by 2030. The need to consider various timescales for AI development was emphasized, as focusing solely on 2035 could overlook important developments on other timescales. It was agreed that consideration of different dates (e.g., 2027, 2030, 2035) would be useful to address the uncertainty around the pace of AI development.

Refine the framework

The "continual learning" column in the framework was discussed, with concerns raised about its relevance. It was noted that, even without continual learning, regular updates of AI models could deliver many similar impacts to continual learning. Agency, autonomy or long-time horizon planning were put forward as a potentially more relevant 3rd category.

Distinguish between levels of advancement
There were questions about how to differentiate between incremental and substantial AI advancements. The economic dynamics of these advancements were also discussed, particularly in terms of investment. Experts noted that a plateau scenario would likely imply a collapse of investment in AI relative to current levels. Experts also highlighted that AI capabilities are likely to diffuse along a series of S-curves, suggesting some continued change even in a capabilities plateau.

Prof. Russell concluded by announcing plans to distribute a survey to gather further input from participants. The results will inform the design of future workshops aimed at refining the scenario framework to ensure meaningful and concrete discussions.

Presentation and discussion on the agentic AI deep dive

Live agentic AI demo

Ashley Zlatinov, Head of Product Public Policy in Anthropic, provided an AI demo demonstrating increasingly capable AI systems and presentation. Ashley introduced the concept of agentic AI, emphasizing that although the term lacks a precise definition, it generally refers to systems that can perceive their environment, use tools, take actions, and operate autonomously for extended periods. Through a demonstration of Claude 3.7 Sonnet, Ashley illustrated how the AI could assist congressional staff in AI policy research and legislative drafting by automating tasks such as analyzing legislative texts, generating policy recommendations, and drafting emails. Looking ahead, Ashley noted that the industry is advancing toward more autonomous AI systems capable of executing complex, multi-step tasks with minimal human intervention. She noted that Anthropic, as a safety-focused company, recognises that AI tools have significant implications for governance, privacy, safety, and security. To address these concerns, she noted they are approaching safety work rigorously and holistically across the product lifecycle and thoughtfully deploying technologies to ensure proper safeguards and responsible deployment.

Discussion among meeting participants

Participants emphasized the importance of human oversight of advanced AI systems, pointing out that AI-generated content or actions should be reviewed at critical points to ensure accuracy and reliability. The issue of hallucinations in multi-step processes was also raised, with concerns about compounding errors and the need for verification at each stage. Additionally, the conversation touched on standardisation and interoperability, stressing the importance of open standards to ensure broad adoption and compatible solutions enabling consumer choice.

In the closing remarks, Yuko Harayama, the Secretary General of GPAI Tokyo Expert Support Center, emphasized the importance of keeping humans in the loop and exploring ways AI can effectively complement human involvement.

Presentation on Agentic AI deep dive

Francesca Rossi presented a deep dive on Agentic AI, outlining the following proposals:

Aim

- Take stock of the various definitions of agentic AI systems.
- Identify examples and evidence of the use of agentic AI systems.

Francesca also shared a presentation outlining a proposed definition of agentic AI, its unique characteristics and associated benefits, challenges and risks. She highlighted that these systems have the agency to act based on explicit or implicit objectives. These systems are composed of various components such as LLMs, tools, symbolic planners, and datasets, which interact with each other, sometimes without full user control. The three key levels of agentic actions are: (1) taking actions that impact the world (e.g., booking a flight or transferring money), (2) consulting resources and using tools, and (3) deciding which processes or resources to select (which could introduce bias or unintended consequences). Francesca noted the amplified risks of agentic AI, including issues with fairness, value alignment, over-reliance, and bias. There are also challenges in evaluating these systems, such as the inability to assess risks at the system level, transparency, reproducibility of behavior, and compliance with regulations. These risks create a complex, open-ended environment that poses significant challenges for managing and understanding agentic AI systems.

Francesca then opened the floor, inviting feedback on the following discussion questions.

- How should agentic AI be defined? Is agentic AI a coherent category and if so, what defines it and delineates it from related concepts such as autonomy/autonomous AI systems?
- What are existing examples of agentic AI systems and how do these map onto potential definitions?

Participants provided feedback and raised several important points:

Defining agentic AI

Some experts suggested that AI systems have always exhibited agency by pursuing objectives, but the addition of new action channels (e.g., sending emails, modifying files) significantly alters their impact. Others emphasized the need to distinguish between agency and autonomy, questioning whether AI truly makes decisions or merely simulates decision-making. One expert also questioned whether "agentic AI" constitutes a coherent category, given that AI systems are engineered through diverse methods and lack a shared evolutionary origin, unlike natural agency.

Establishing liability and responsibility in agentic AI

Concerns were raised about AI disregarding human instructions and the liability issues arising from agentic AI's actions—whether responsibility should fall on users or developers.

Assessing the role of communication and initiative

Beyond autonomy, the importance of communication capabilities was highlighted, as agentic AI systems need to effectively interact with other entities to be functional.

Francesca acknowledged the interesting points raised during the discussion, highlighting various perspectives such as the nature of agency, the nature of autonomy, taking initiative, having goals, etc. She mentioned that the sub-groups will explore existing definitions and distinctions of agentic AI, comparing them with past AI concepts discussed by experts. Francesca thanked everyone for their input, assured that the next steps will be shared soon, and confirmed that the ideas presented had been noted.