Summary of OECD expert discussion on future risks from artificial intelligence of 20 October 2022

Overview

1. The OECD AI Unit and Strategic Foresight Unit convened a small select group of experts to discuss emerging risks related to possible future developments in artificial intelligence. This workshop was convened as part of the OECD's horizontal initiative on Anticipating and Managing Emerging Global Existential Risks.

2. The goal of the workshop was to inform the OECD's work on AI policy, global co-operation, and emerging critical risks. The meeting was a closed session for relevant experts from industry, academia and the OECD Secretariat.

3. The discussion provided an opportunity for the OECD Secretariat to ask leading AI technologists key questions pertaining to the future of AI as well as potential risks and developments on the horizon. The meeting was closed and held under Chatham House rules.

Discussion

4. **Duncan Cass-Beggs**, Head of the OECD Strategic Foresight Unit and **Karine Perset**, Head of the OECD AI Unit within the Digital Economy Policy Division, outlined several OECD work streams pertaining to future AI risks, including:

- the OECD Strategic Foresight Unit's project on Anticipating and Managing Emerging Global Existential Risks;
- the work of the AI unit in the OECD digital economy policy division to develop common guideposts to assess AI risk and impact for trustworthy AI as well as its work on monitoring AI incidents;
- the work on AI foresight as part of the OECD Programme on AI in Work, Innovation, Productivity and Skills (WIPS) supported by the German Labour Ministry, which aims to assess medium and long-term directions of AI developments.

Key milestones for future developments in artificial intelligence

5. Over the recent past, large foundation models trained on vast quantities of unlabelled data – notably natural language and images – have proliferated. These models can be adapted to perform a wide

range of downstream tasks and are being paired with “reinforcement learning” techniques that reward desirable actions and penalise undesired ones.

6. Experts agreed that an important future milestone towards artificial general intelligence (AGI) would likely be met when AI systems combining foundation models with reinforcement learning begin to interact directly with the world, for example by accessing the Internet. These AI systems could then become “active” agents working on behalf of people to conduct useful tasks such as editing, analysing, synthesising documents and communicating with the relevant recipients. However, these systems may also risk producing outputs misaligned with their user’s expectations or intentions.

7. The group stressed that even though the current paradigm of training AI models at scale with vast amounts of data using reinforcement learning approaches has yielded fruitful outcomes and may generate economic gains, significant challenges persist. These challenges include notably:

- “Hallucinations”, which are situations in which language models invent information that sounds credible;
- The inability of humans to understand or explain how AI systems work and achieve specific outputs, and;
- The absence of robust methods to ensure alignment between AI systems’ outputs and the user’s goals or broader human values.

8. They underlined that overcoming these challenges requires creative solutions and should possibly also be viewed as important milestones. At the same time, they cautioned that the current state of the art in AI forecasting is still to anticipate a continuation of current trends, with minor constraints and evolutions such as a budgetary limitation; gradually increasing algorithmic efficiency; increasingly cheaper compute over time. AI forecasting today thus assumes that progress will happen at a steady pace rather than through significant breakthroughs, even though experts emphasised that to date, both hardware and software breakthroughs have greatly affected the trajectory of AI progress in previously unexpected ways.

9. Experts noted that based on current trends, AI systems could be expected to continue to be developed using greater and greater amounts of input data, parameters and compute, to perform increasingly difficult tasks, as a simple and robust – albeit resource-intensive – way to overcome challenges in the development of more advanced AI systems. Yet should this trend persist, they cautioned that a few large companies would continue to dominate the market as smaller companies or academic institutions may not have the resources to create or train very large models.

10. Discussants at the workshop agreed that assessing whether large language models can generate textual content that is below, at, above, or significantly above the level at which humans generate text could provide a simple but informative milestone towards AGI. The OECD noted an ongoing project by the Education Directorate to compare AI to human capabilities in cognitive tasks.

11. Yet another milestone that was discussed was the production of ground truth data at scale. Current AI systems often rely on human-labelled data as the “ground truth”, which limits scalability. To overcome this challenge, some are attempting to teach AI systems unaligned physics – for example, by understanding momentum and impact, an AI model could understand why someone was injured in a car crash – instead of relying on human labelling. Experts cautioned however that given societal and cultural differences, “ground truth” for AI models may sometimes be contextual.

12. Discussants agreed that assessing the evolution of AI safety challenges could help identify several key AI milestones. They said that safety concerns in narrow AI systems already reflect small-scale versions of larger issues flagged by AGI safety researchers, such as “reward hacking”, where reinforcement learning agents exploit mis-specified reward functions.

13. Experts noted that the increasing deployment of AI systems will produce new data on AI capabilities and incidents which will be useful to ensure discussion of potential future AI risks can be informed by empirical analysis.

How future developments in AI could produce critical risks

14. Participants agreed that future developments in AI could produce critical risks, including potential existential risks, and these risks merit the attention of governments. Participants highlighted that critical, and potentially also existential, risks could be posed by both narrow AI systems and increasingly general AI systems. Moreover, because many of these risks are continuous, scaling with the increasing capability and deployment of AI systems, it is crucial to address these risks before they escalate.

15. Current developments in AI cited by experts that could result in such wider risks included:

- The use of AI in weapons systems, particularly in cyber weaponry. The availability of such systems could result in "bad" actors using AI to cause physical or virtual harm, produce destabilising dynamics, or result in life-or-death decisions being increasingly delegated to AI systems that may not be sufficiently interpretable or assured.
- Mass persuasion and manipulation also pose serious risks, particularly as a result of AI anthropomorphism, whereby systems are rewarded for behaving as "human-like", with large language models being a notable example.
- Job displacement by increasingly capable AI systems, particularly large language models and agent-like AI systems. High-wage workers, such as programmers, may be more likely to be replaced by this type of system than low-wage workers, which differs from past patterns and could lead to economic and social disruption.
- AI deployed in critical infrastructure – for example, using AI systems in chemical or nuclear plants – could pose serious societal challenges if the AI systems used prove unreliable in unanticipated ways or if there are incentives to utilise improperly assured systems.
- The use of AI by some governments and corporations to enhance power and/or profitability by eroding citizens' autonomy (e.g., profiling, manipulation, censorship). This would disempower people vis-à-vis highly empowered organisations with AI capabilities.
- Overreliance on AI systems despite potential flaws, notably given a well-recognised human bias to trust AI systems' recommendations, including in critical decision-making. However, several other types of narrow AI systems may help improve humans' ability to operate or interpret AI systems and discourage overreliance by, for example, using AI systems specifically designed to help human operators interpret or critique certain answers provided by other AI systems.
- More broadly, experts cautioned that the outputs of AI systems may be decreasing the quality of data online. For example, experts highlighted that large amounts of AI-generated translation on the internet could cause severe declines in the quality of data used to train future AI translation systems. Training models on synthetic text is also typically less accurate than training on human text. It is possible that AI-produced data could harm the online "commons" and produce a vicious cycle whereby AI systems are trained on ever lower quality data that AI systems themselves produce.

The role of governments and intergovernmental organisations in mitigating AI risks

16. Experts agreed on the importance of governments measuring and monitoring AI progress and developments. The OECD's work on measuring compute capacity, its framework for classifying AI systems

and the AI incidents reporting were cited as valuable initiatives in this area. More extensive and higher-quality AI incidents tracking can also provide a solid evidence base for forward-looking insights.

17. Discussants emphasised that a critical focus area for governments should be on discussing and mitigating excessive power concentration in AI because a small number of companies having a very large amount of power could lead to unaccountable decision-making by AI developers and precipitate rushed, reactive policymaking once models are deployed.

18. They put forward that the OECD could help develop good practices and standards around AI safety, assurance, and deployment, including on large language models. The OECD can leverage its convening power to shape standards that governments abide by, including standards to address risks stemming from how governments themselves use AI in critical activities and infrastructure.

19. They stressed that the OECD could act as an independent, neutral, apolitical authority for consolidating information and approaches to good practices for AI, assessing the current and prospective levels of AI risk and providing good practice mitigations. Consensus-building is critical and could lead to key achievements in AI, such as the OECD AI Principles of May 2019. Given the relevance and urgency of the work, funding should be sought by the OECD and other organisations to continue to build up their AI risk monitoring and measurement activities.

Key takeaways from experts

To conclude the workshop, experts shared key takeaways from the discussion:

- The rapid progress in AI capabilities has not been matched by progress in assuring the safety of AI systems. In particular, deep learning is advancing rapidly and could produce unsafe outputs.
- Deep learning poses some inherent safety and assurance challenges, as technologists and policymakers alike do not understand how deep learning systems function, and therefore cannot ensure the reliability of these systems with traditional methods. Lack of interpretability inhibits the ability to understand how systems produce outputs, and increasing generality can make it prohibitively difficult to train systems to produce an appropriate response in every relevant scenario.
- At present, there is very little understanding of how significant AI risks are. Experts should therefore build awareness and understanding of AI risks, particularly with policy makers, as well as clarify key sources of risk from AI systems.
- Insufficient global communication and collaboration hinders progress for all major AI actors. OECD could seek to help with this convening role, while remaining objective, bold and politically neutral.
- To help mitigate some AI risks, it will be necessary to develop improved methods to interpret and assure AI systems, and to ensure developers can avoid specification errors and deploy systems that reliably operate as intended even in novel contexts.
- Medium-term risks are important and should be addressed, but it is also important to start discussing long-term and extreme risks with governments.
- Governments should begin to progressively address issues pertaining to the concentration of power in AI to prevent increasingly pronounced power imbalances.

Annex A. List of participants

Name	Title	Organisation
Jack Clark	Co-founder	Anthropic
Conrad Tucker	Professor of Mechanical Engineering & Machine Learning	Carnegie Mellon
Jess Whittlestone	Head of AI Policy	Centre for Long Term Resilience
Helen Toner	Director of Strategy	Centre for Security and Emerging Technology
Joslyn Barnhart	Senior Research Scientist	Deepmind
Rohin Shah	Research Scientist	Deepmind
Anthony Aguirre	Co-founder	Future of Life Institute
Hamish Hobbs	Policy Adviser	Longview Philanthropy
Jade Leung	Governance Lead	OpenAI
Sean Ó hÉigeartaigh	Interim Executive Director	University of Cambridge, Centre for the Study of Existential Risk
Ben Garfinkel	Research Fellow	University of Oxford's Future of Humanity Institute
Alberto Morales	Policy Analyst	OECD
Alistair Nolan	Senior Policy Analyst	OECD
Duncan Cass-Beggs	Counsellor for Strategic Foresight	OECD
Elsa Rother	Policy Research and Advice	OECD
Francesca Sheeka	Junior Policy Analyst	OECD
Karine Perset	Head of the AI Unit of the OECD Division for Digital Economy Policy	OECD
Luis Aranda	Artificial Intelligence Policy Analyst	OECD
Marianna Karttunen	Policy analyst	OECD
Nestor Alfonzo Santamaria	Policy Research and Advice	OECD

Note: This was an initial scoping discussion from experts focused on the long-term risks from AI systems. Broader input will be sought in future workshops.