OECD Expert Group on AI Futures – Meeting 3 (9 November 2023)

Background

The OECD.AI Policy Observatory and OECD Strategic Foresight Unit convened the third meeting of the Expert Group on Artificial Intelligence (AI) Futures on 9 November 2023. 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. The list of participants for this third meeting can be found in the Participants section towards the end of this document.

Introduction and context

The hybrid meeting was part of full-day set of sessions including other OECD.AI Network of Experts (ONE AI) meetings, open to members of ONE AI, delegates from the OECD Working Party on AI Governance (AIGO), and virtually to the public. A recording was made of the meeting and is available on YouTube. Usually, expert group meetings are held under the Chatham House Rule, but this time, in light of the pertinence of the topics discussed, the meeting was held openly in order to promote transparency and awareness.

Karine Perset, Head of the OECD.AI Policy Observatory, kicked off the meeting with an introduction to the session's scenario exploration exercise, before the three co-chairs introduced themselves. The co-chairs explained that the scenarios were developed based on past expert group discussion and reminded the audience that the scenarios were not predictions, but plausible cognitive stretching exercises.

Representing the OECD Strategic Foresight Unit, Hamish Hobbs and Dexter Docherty first provided a recap on the themes for desirable and undesirable futures discussed in the last expert group session, then presented three scenarios (see below) to be discussed and provided guiding questions on how they could be approached. The co-chairs facilitated discussion on the three scenarios, allowing approximately three minutes for each of the 27 members intervening in the session.

Scenario 1: Benevolent Oligopoly

Box 1. Context for Scenario 1 – Benevolent Oligopoly

AI capabilities and global AI safety overseen by a handful of companies in advanced economies

The size and complexity of AI models continue to rise, resulting in both increasingly capable AI models and increasing development costs. The costs to train new models, adhere to regulatory requirements and acquire scare AI talent have become so high that only five companies are able to operate at the cutting edge.

High costs (e.g. development, regulatory compliance and liability vis-à-vis legal battles related to AI) mean leading firms carefully guard their intellectual property. The dominant firms offer access to their technologies as a paid premium service that has demonstrated clear value to those who can afford it. These services have improved the productivity of the largest firms and many SMEs. Leading AI development firms have defended tight control of their models on safety grounds, citing evidence that guardrails placed on foundation models often fail if the models are modified or can be overridden by malicious actors. The AI start-up ecosystem lags behind as the dominant AI firms have found a way to secure financing, attract talent and comply with oversight, privacy and safety requirements with greater success than AI start-ups.

The gap between the dominant firms and other providers, especially those outside of OECD countries, is growing. Large non-OECD economies have been able to procure services from the major firms that meet their domestic or develop sufficient alternative programmes themselves. Smaller and developing countries are disadvantaged. First, they have less access to cutting-edge AI catered to their specific needs. Second, to the extent they do have access, they have little option but to accept these models and any associated rules even though they had no hand in shaping them.

Discussion questions:

- What risks and benefits from AI might you expect to materialise in this scenario?
- How can governments ensure sufficient democratic oversight and distribution of benefits of AI?

Expert group co-chair Francesca Rossi moderated the exploration exercise for this scenario. Interventions were made by expert group members Sarah Myers West, Helen Toner, Charles Fadel, Carolyn Ashurst, Conrad Tucker, Azeem Azhar, Yoshua Bengio, Aaron Maniam, Juraj Čorba, Dan Faggella, and Emmanuel Kahembwe (in chronological order; see Annex for titles, organisations, and hyperlinks to bios).

Full individual interventions can be viewed on the public YouTube video. The summary below identifies some over themes from the discussion, with key members discussing the theme indicated in parenthesis.

Assumptions

On this scenario, some posited that the recent trend of open sourcing might help mitigate this oligopolistic market from materialising (Charles Fadel), along with recent reduction of compute necessary due to more efficient learning models (Toby Walsh). Meanwhile, several experts (Carolyn Ashurst, Helen Toner, Sarah Myers West) believed that this model is close to the existing state of affairs, in part due to current industry requirements demanding resources which only large technological firms can afford, exacerbated by governments using public funds to support dominant firms (Sarah Myers West).

Potential benefits

According to experts, benefits of this market structure would include **ease of regulation**, where it would be more straightforward to coordinate a shared set of rules and implement red-lines and lessen incentives to cut corners on safety (Dan Faggella, Helen Toner).

Additionally, this scenario could help to foster **interoperability**, standardising technological education globally and increase geographical job mobility (Conrad Tucker).

Potential risks

On the other hand, the dominance of firms could allow a **race to the bottom**, with weak scrutiny and testing (Sarah Myers West), similar in the way social media has developed (Charles Fadel). Importantly, many experts pointed out our dependence on individual firms to be especially dangerous, with only the personality of people in charge of them defining their level of benevolence (Charles Fadel).

Several experts discussed that outsize power concentration and influence can perpetuate significant **inequality** (Aaron Maniam, Carolyn Ashurst, Sarah Myers West). With foundational AI models possibly serving as the infrastructural base for many future services, those with political power or influence could see much potential in collaborating with technological leaders to give to the people the illusion of democratic control even if it is more self-serving than citizen-serving (Azeem Azhar). Large companies can thus harness their power and influence to shape global narratives and potentially impact issues as critical as democracy, meaningful consent, human autonomy, mental well-being, and interpersonal relationships (Carolyn Ashurst).

For instance, the Global South cannot shape oversight of the software they also cannot afford to access which however has been document to rely on their labour for end-uses such as content moderation (Carolyn Ashurst, Helen Toner), seen by some as a form of (neo)feudalism (Emmanuel Kahembwe).

With core corporate interests coming to the fore, there is also the risk of a proliferation of **empty commercial applications**, where platforms chase user engagement (Helen Toner) and even escapism into AI-enabled virtual (AR/VR) ecosystems, eroding human agency and rendering us obsolete (Dan Faggella).

Potential solutions and approaches

To address these issues, many experts agree that **strong regulations** are needed, especially as AI becomes increasingly capable. Such regulations, experts suggested, should be driven by governments and civil society-led, citizen-oriented, transparent, and democratic **institutions** that can levy significant penalties if warranted, rather than large technological corporations controlling their own oversight (Aaron Maniam, Azeem Azhar, Carolyn Ashurst, , Emmanuel Kahembwe, Yoshua Bengio,). This can be seen as rethinking of AI as a public commons instead of a private service to, with civil society, construct a competitive environment for transparency (Azeem Azhar).

Experts have multiple recommendations for these institutions—delineating between foundational and applied approaches (Azeem Azhar), harnessing financial markets systemic risks analysis (Juraj Čorba), improving access and oversight and having a hand in the distribution of benefits and taxing companies to subsidise safety research (Helen Toner). With the rapid pace of developments, some experts stress the need to act swiftly (Charles Fadel).

They also generally agree that **geopolitical risks** need to be mitigated, in part by enhancing global cooperation and engaging meaningfully with countries that are not always engaged in international AI cooperation, such as China (Aaron Maniam, Helen Toner, Juraj Čorba).

Another key solution stressed was to develop **human capacity**, by investing in (student) exchanges and knowledge transfers (Conrad Tucker), individual skillsets and education (Aaron Maniam), including AI courses for government employees (Emmanuel Kahembwe).

Ultimately, some have asked for us to decide on what future we hope to achieve—from preserving to completely revolutionizing our current way of life (Dan Faggella).

Scenario 2: Democratised benefits and distributed risks

Box 2. Context for Scenario 2 - Democratised benefits and distributed risks

Open-source models lead to mixed impacts

Open-source proves to be the best business model for cutting-edge AI. Start-ups from around the world drive progress with applications based on open-source foundation models or fine-tuned versions of these models that facilitate catered solutions in specific use cases and contexts. The barriers to entry for deploying AI are low due to a vibrant open-source community that provides peer support in creating both free and paid beneficial AI services. Large corporations still develop proprietary models, but consumers are able to choose between these and open-source models, mitigating corporate concentration risks. Some progress has been made on alignment and ethical oversight challenges in specific contexts through best practices identified and promoted by the open-source community.

However, misuse risks have begun to materialise as open-source AI tools are also available for bad actors. Personalised disinformation campaigns crop up during election periods posing a threat to democratic functioning. Scammers use anthropomorphised AI systems to spoof identities to extract sensitive information and commit fraud or blackmail. SMEs, large firms and governments are suffering from more frequent AI-enabled cyberattacks—though the open-source community has driven innovation in cyber-defence. Security agencies are concerned after intelligence reports find that a terrorist group is using a combination of AI tools and desktop DNA synthesis machines in an attempt to design lethal new pathogens. Experts warn that open-source AI tools combined with advances in DNA synthesis are dramatically lowering barriers to access for biological weapons.

Discussion questions:

- What risks and benefits from AI might you expect to materialise in this scenario?
- How can governments best manage the risks in this scenario?

Expert group co-chair Stuart Russell moderated the exploration exercise for this scenario. Interventions for this scenario were made by expert group members Rebecca Finlay, Sebastian Hallensleben, Amir Banifatemi, Clara Neppel, Benoit Bergeret, Marko Grobelnik, Graham Taylor, and Joscha Bach (in chronological order; see Annex for titles, organisations, and hyperlinks to bios).

Assumptions

Some experts point out that Open-Source AI is a paradigm shift which has already started to happen (Graham Taylor, Marko Grobelnik), yet in a **different manner** than what we may assume. Clara Neppel explained that foundation models are pre-trained, in comparison to normal code, and that feedback mechanisms work differently, and new techniques would need to be invented to deal with issues such as alignment. Benoit Bergeret also warned that ground-up, community-driven open-source innovation mainly currently applies on the usage-level while impactful deep innovation, such as deep learning or the transformer architecture driving leading-edge systems and models, is still dominated by private firms.

Potential benefits

Experts generally agreed that by decentralising, democratising, and thus expanding the access to AI development to academics and regardless of economic status or geography, open sourcing can widen and transform the competitive landscape for **innovation** to pool resources in the spirit of coopetition to drive progress on global issues, from climate change to healthcare (Amir Banifatemi, Clara Neppel, Graham Taylor, Joscha Bach, Marko Grobelnik, Rebecca Finlay, Sebastian Hallensleben, Toby Walsh). In addition, open-source models could be more resilient, not relying on a single entity to maintain funding and support for model hosting and development, thus inspiring trust in their userbase regarding the continuity of AI services (Sebastian Hallensleben). Linux was cited as a past example of open-source software allowing community developers to pool resources to finetune and foster downstream applications as part of a broader ecosystem (Clara Neppel, Graham Taylor, Sebastian Hallensleben). As part of this, civil society can also be part of the evaluation and auditing process to increase transparency (Rebecca Finlay).

Despite short term disruption and job loss, some experts view open-source AI to transform productivity and reap long-term benefits (Amir Banifatemi, Joscha Bach).

Potential risks

However, experts also recognised the difficulty in regulating and controlling AI once open-source models are released (Amir Banifatemi, Sebastian Hallensleben) leading to potential anarchy in AI uses (Marko Grobelnik), and harms can materialise through malicious actors, which could use the freely available technology for innovating in weapons or bioterrorism (Clara Neppel, Joscha Bach, Rebecca Finlay).

In a broader sense, Benoit Bergeret also noted the potential backlash against AI and discontent that could arise from job losses, possibly leading to shifting more to a more populist political background. If this occurs using open source models, it could push people into embracing more proprietary models, reinforcing the power concentration challenges discussed in Scenario 1.

Potential solutions

Experts primarily stressed the importance of practical and technical solutions, from safety mechanisms such as early warning systems and systems that can fail safely, to government oversight, liability laws, standards testing, and penalties for misuse (Amir Banifatemi, Clara Neppel, Rebecca Finlay). Some also discussed a preference towards more targeted monitoring and restrictions on specific misuse such as human impersonation and addictive algorithms through risk-assessment methods rather than broad sweeping legislation (Graham Taylor, Joscha Bach).

In addition, experts also discussed the need to provide soft solutions, from strengthening trust and digital space resilience (Clara Neppel) such as misinformation monitoring (Joscha Bach) and new models of content curation (rather than data harvesting and paid advertisements) (Sebastian Hallensleben) to investing in workforce development and providing incentives and resources to help open source communities engaging in public-serving AI research and development embrace responsibility (Amir Banifatemi, Benoit Bergeret).

Scenario 3: Pulling the plug

Box 3. Context for Scenario 3 - Pulling the plug

Regulator calls for implementing a global moratorium on further AI development

OECD countries agree to implement a strong risk-based AI oversight regime for leading-edge AI. Training of high-impact AI systems in OECD countries is controlled via a licensing regime, which requires AI developers to demonstrate that their AI systems will not pose unacceptable safety, ethical or privacy risks. Highly generalised AI services developed by large firms and innovative start-ups are integrated into core decision-making functions of leading firms and governments with great success. Large non-OECD economies (e.g. BRICS+) are not members of the oversight regime that has been adopted by OECD countries. The AI capabilities of these large non-OECD economies are behind but near the cutting edge for most AI tools, including being world leading in some specific AI techniques and applications.

A regulator in a large OECD country determines that a critical safety threshold could be surpassed and prohibits an upcoming training run by a leading AI lab. The regulator announces that it is unsafe for leading labs to continue developing more advanced general AI systems. The regulator believes these systems are approaching human levels of general intelligence, that insufficient safeguards are in place, and that there is a risk of loss of human control. OECD leaders are briefed and advised that it may be necessary to cease the development of more advanced general AI systems indefinitely, until safety can be assured. How this will be achieved is currently unclear—especially given no fully global agreement on frameworks for AGI governance exist.

Discussion questions:

- What might be the criteria for pulling the plug on AI development?
- What is the system needed to manage loss of control risks?
    - What systems would be needed to implement an order to stop AI development?
    - Who would need to be involved in making such a decision?
    - What would be necessary to ensure sufficient global agreement on pausing AI development?

Expert group co-chair Michael Schönstein moderated the exploration exercise for this scenario. Interventions for this scenario were made by expert group members Stuart Russell, Yoshua Bengio, Toby Ord, Sean Ó hÉigeartaigh, Markus Anderljung, Duncan Cass-Beggs, Ziv Katzir, and Pam Dixon (in chronological order; see Annex for titles, organisations, and hyperlinks to bios).

Where

Despite, the upside of being able to come to an agreement to pause AI development in the name of human safety, experts caution about the **difficulty in knowing and determining** when we reach a critical safety threshold (Sean Ó hÉigeartaigh, Stuart Russell), with Stuart Russell nothing in particular that we should never let it get to this point, and that the putting in place processes for ensuring human control over machine sis critical or such recognition may already be too late. According to several experts, control risks are not binary, and it is risky to use data or compute measurements such as FLOPS and parameters as yardsticks (Duncan Cass-Beggs, Markus Anderljung, Ziv Katzir) since even smaller models or spin off applications can produce harms (Pam Dixon, Stuart Russell). Several experts agree that broadly encapsulating **capability-based milestones** are more optimal measurements, including autonomy, strategic planning, ability to deceive, and self-replication, or specific risk outcomes such as an ability to launch large scale cyberattacks (Duncan Cass-Beggs, Sean Ó hÉigeartaigh).

Who

Conceptually, experts generally agree on putting **burden of proof on developers**, who must analytically demonstrate system safety to regulators before deployment, such as in fields such as pharmaceuticals, nuclear, aviation, and infrastructure (Duncan Cass-Beggs, Stuart Russell, Yoshua Bengio). However, many also emphasise the importance on having **regulation proportionate** to their risk posed, through risk classification methods (Duncan Cass-Beggs). Examples of models proposed include company-led Responsible Scaling Policies, which incorporate appropriate safeguards at structured pause points (Toby Ord). Thus, frontier model and bad-actor misuse risks require most oversight (Duncan Cass-Beggs, Ziv Katzir), along with instances where models are given the most agency (Toby Walsh).

Nonetheless, Duncan Cass-Beggs also pointed out that current decisions that impact numerous lives around the world are already made by a few multinational corporations, thus, despite the emergence of company-proposed regulation, perhaps the global community needs to step up to develop, implement, and enforce a robust decision making system.

When

Even though some experts think we need to be concerned now (Duncan Cass-Beggs), they emphasise the importance of **finding the right moment** to make the decision to pause AI development, with overly cautious premature calls eroding the value of the decision, and disorderly late pause possibly leading to a catastrophe (Sean Ó hÉigeartaigh, Toby Ord). However, complicating matters is that even if such a pause is made just in time, we would not have a buffer to prevent bad actors to take the final step.

Furthermore, experts warn that a pause is only a **temporary solution**, since implementing computational limits will eventually allow other actors to catch up, algorithms could become more efficient, or smaller models might be able to cause significant harms, and thus this fix is merely a tool to buy time to learn to develop safer systems and increase societal resilience (Markus Anderljung, Ziv Katzir)

What

They also emphasise the difficulty of realising this pause. **Necessary preconditions** include consensus on the necessity, and thus political will for the pause, concentrated compute supply chains, large frontier systems compute running requirements, detection mechanisms, and mandated off switches (for both open and closed source systems) (Markus Anderljung, Sean Ó hÉigeartaigh, Stuart Russell). However, the ability to control and turn off open source systems was debated, with Yoshua Bengio expressing doubt on the possibility of ensuring open-source systems safety since protections can be removed and cannot be regulated once access is given irreversibly to bad actors, though sandboxes and limited access to researchers could ameliorate this, which may move more towards open source alternatives rather than full open source in ways that still yield democratic benefits.

How

To ensure that these preconditions for a pause are met, oversight bodies can be fostered, collaborating with leading researchers and companies, and boosted by inclusive international-level **multilateral agreements** to impose enforced economic penalties and cheating countermeasures (Pam Dixon, Sean Ó hÉigeartaigh, Yoshua Bengio), with specific efforts to include state actors who are now currently part of international norms like the OECD AI Principles (Yoshua Bengio) though Pam Dixon called for modesty in our regulatory capabilities.

Annex – List of Participants

Markus Anderljung - Head of Policy at the Centre for the Governance of AI.

Carolyn Ashurst - Turing Research Fellow at The Alan Turing Institute.

Azeem Azhar - Founder of Exponential View.

Joscha Bach - Principal AI Engineer at Intel Labs.

Amir Banifatemi - Co-founder of AI Commons.

Yoshua Bengio - Professor at University of Montreal and Founder and Scientific Director at MILA, Quebec AI Institute.

Duncan Cass-Beggs - Executive Director of the Global AI Risk Initiative at the Centre for International Governance Innovation (CIGI).

Benoît Bergeret - Executive Director at Metalab, ESSEC Business School.

Jamie Berryhill – AI Policy Analyst at OECD.

Juraj Čorba - Senior expert, Digital Regulation & Governance - Slovak Ministry of Investments, Regional Development and Informatization.

Pam Dixon - Founder and Executive Director at World Privacy Forum.

Dexter Docherty – Strategic Foresight Analyst at OECD.

Charles Fadel - Founder & Chairman at Center for Curriculum Redesign.

Daniel Faggella - Head of Research, CEO at Emerj AI Research.

Rebecca Finlay - CEO at Partnership on AI.

Marko Grobelnik - AI Researcher & Digital Champion at the AI Lab of Slovenia's Jozef Stefan Institute.

Sebastian Hallensleben - Head of Digitalisation and AI at VDE Association for Electrical, Electronic & Information Technologies.

Hamish Hobbs- Policy Advisor to the OECD Strategic Foresight Unit.

Emmanuel Kahembwe - CEO - VDE UK.

Holden Karnofsky - Director of AI Strategy at Open Philanthropy (provided written responses).

Ziv Katzir - Head of the National Plan for Artificial Intelligence Infrastructure at Israel's Innovation authority.

Aaron Maniam - Fellow of Practice and Director, Digital Transformation Education at Blavatnik School of Government, University of Oxford.

Sarah Myers-West - Managing Director of the AI Now Institute.

Clara Neppel - Senior Director at IEEE European Business Operations.

Sean Ó hÉigeartaigh - Interim Executive Director of CSER (Centre for the Study of Existential Risk) at University of Cambridge.

Toby Ord - Senior Research Fellow at University of Oxford.

Karine Perset - Head of OECD.AI Policy Observatory at OECD

Francesca Rossi - IBM Fellow and AI Ethics Global Leader.

Stuart Russell - Professor of Computer Science at the University of California, Berkeley and Director of the Centre for Human-Compatible Artificial Intelligence.

Michael Schönstein - Head of General Digital Policy - Federal Chancellery of Germany.

Graham Taylor - Research Director - Vector Institute for Artificial Intelligence.

Helen Toner - Director of Strategy and Foundational Research Grants at Center for Security and Emerging Technology (CSET).

Conrad Tucker - Professor of Mechanical Engineering at Carnegie Mellon University.

Toby Walsh - Chief Scientist of UNSW.AI (provided written responses).