OECD Expert Group on AI Futures – Meeting 2 (18th & 20th September 2023)

Background

The OECD.AI Policy Observatory and OECD Strategic Foresight Unit convened the second meeting of the Expert Group on Artificial Intelligence (AI) Futures on 18 and 20 September 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 second meeting can be found in Annex A of this document.

Introduction and context

To help accommodate different time zones and enhance inclusiveness, the meeting was divided in two sessions, each with the same agenda (18 and 20 September 2023). The meetings were held under the Chatham House Rule. Karine Perset, head of the OECD.AI Policy Observatory, kicked off the meetings with a general overview and context regarding the format of the double session and its contents. Ms. Perset invited experts to join a hybrid meeting of the expert group on November, along with other OECD.AI Network of Experts meetings of the OECD expert groups on AI Incidents and AI Risks and Accountability (see subsequently published video of the 9 November session).

The discussion of the meeting largely focused on expert group members' thoughts on the early results of an OECD survey about potential future AI benefits, risks and solutions. The expert group had previously discussed the survey concept at the first Expert Group Meeting and had the opportunity to provide feedback on the draft survey items through an asynchronous brainstorming and validation exercise, which resulted in 131 potential future benefits, risks and solutions that subsequently comprised the survey contents. The OECD administered the survey starting in August 2023.

In turn, Jamie Berryhill (OECD.AI Policy Observatory), Hamish Hobbs (OECD Strategic Foresight Unit) and Alistair Nolan (OECD's Directorate for Science, Technology and Innovation) presented insights to lead the discussion. Mr. Berryhill presented the main results of the survey, Mr. Hobbs then further elaborated on how the results will be used to create foresight scenarios, and finally, Mr. Nolan presented on a 2024 workshop to be held as part of the OECD programme on AI in Work, Innovation, Productivity, and Skills (AI-WIPS).

The co-chairs facilitated the discussion items listed below.

Topic 1: Discussion of initial survey results

OECD Secretariat presentation

Jamie Berryhill, AI Policy Analyst from the OECD.AI Policy Observatory, presented the results of the survey. The survey items consisted of 131 future considerations (43 risks, 21 benefits, and 67 solutions) that were identified through extensive OECD research, discussions held in previous events and roundtables, and through the incorporation of expert group member feedback from previous sessions and a follow-up validation exercise.

Each consideration was assessed along two axes: importance and actionability, on a scale from 0-10. A description of how “importance” and “actionability” can be found in Box 1.

Box 1. Context for “importance” and “actionability” provided to expert group members

- **Importance**. In your opinion, how important is it that governments focus on this item? Your thinking should weigh both the magnitude of potential impacts from a given risk, benefit, or policy solution and the probability of these impacts. In assessing impacts you could, for instance, take into account the level of harm that you perceive for potential risks, the societal or economic good that could be yielded by potential benefits, or the magnitude of positive change that could be brought about by implementing potential solutions. Please use the following examples to give a sense of the rating scale:
  - **0**. This is in no way important for governments and the international community.
  - **2.**This is an issue of marginal importance for governments and the international community.
  - **4**. This is a somewhat important issue.
  - **6**. This is an important issue, but not among the most important.
  - **8**. This is among the most important issues for governments and the international community.
  - **10**. This is the most important issue for governments and the international community.
- **Actionability**. Assuming that political will exists, based on all of the factors that you can think of (e.g., feasibility, level of complexity, ease of implementation, current and perceived future technical ability and financial resources, etc.), what is your opinion on how actionable the item is in terms of the ability of a group of like-minded countries to make a significant impact with regards to mitigating potential risks, yielding potential benefits, and putting in place potential solutions. For this item, it may be useful to ground your thinking in the medium-term (over the next 10-20 years).
  - A rating of **zero (0)** implies that you think there is no meaningful way for governments to mitigate this risk, contribute to seizing this opportunity or effectively implement this solution, even through collective action.
  - Ratings in the middle – **five (5)** – imply that a group of like-minded governments could have some agency in mitigating a risk, seizing an opportunity or effectively implementing solutions, but that their success may be partial, uncertain or require an unusually large commitment of resources or high level of global collaboration.
  - A rating of **ten (10)** implies near certainty of almost entirely mitigating the risk, realizing the benefits, or effectively implementing solutions based solely on the actions of like-minded governments.

Given the diversity of the group’s membership, the level of agreement and disagreement is also an important dimension to take into account when understanding views on the survey elements. Thus, a measure of variance (i.e., level of disagreement among members) was also calculated for each item (see Figure 2). Identifying considerations with the highest levels of variance may, for instance, surface areas ripe for discussion and productive debate.

Based on importance and actionability scores, items were graphically presented as belonging to one of four quadrants of a cartesian graph, as in Figure 1 below. Figure 2 illustrates how the dimension of variance could also be visualised.

Figure 1. Responses explored two dimensions

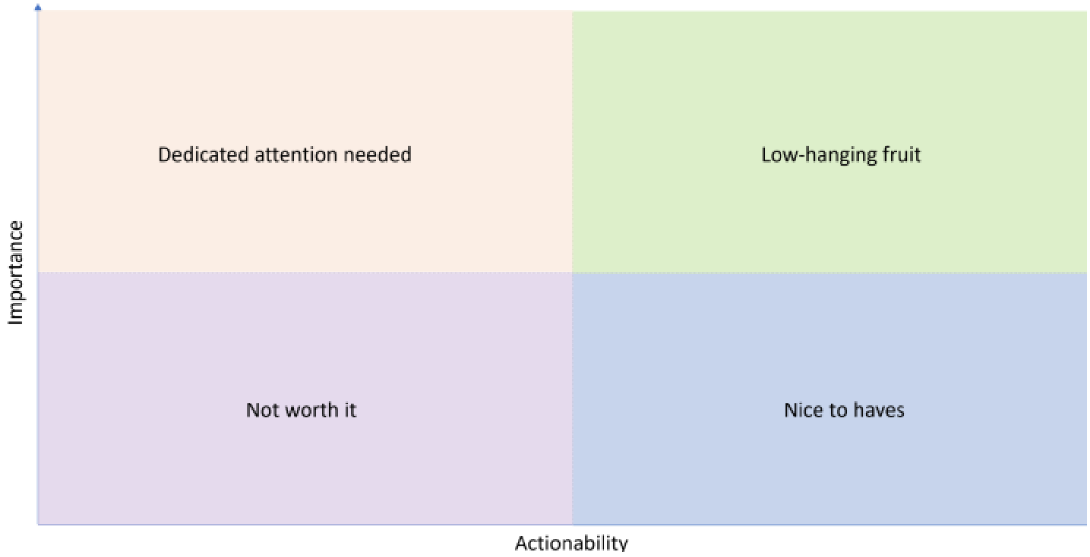
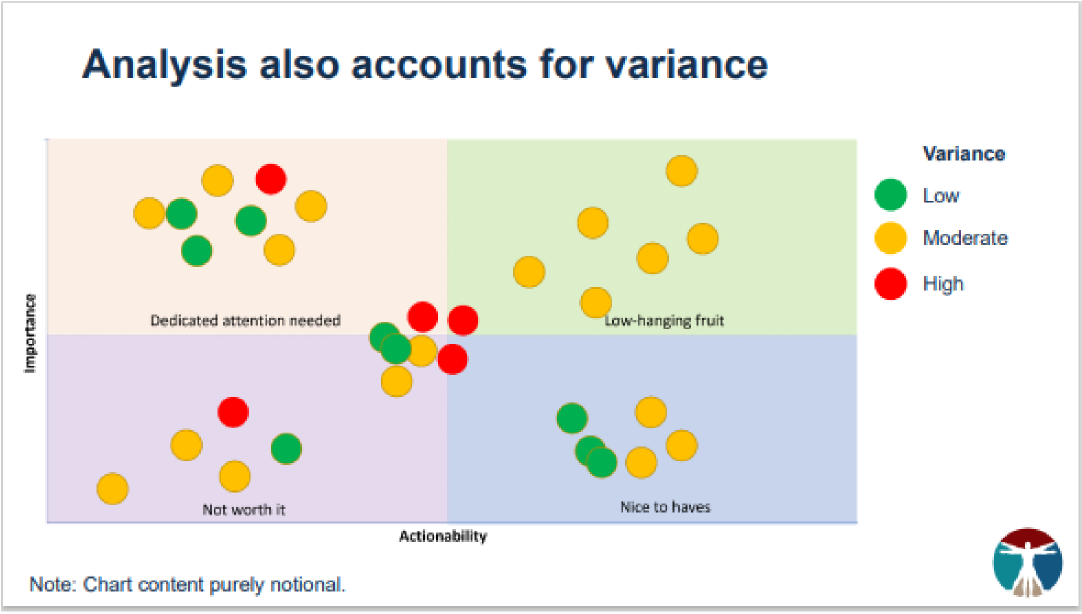


Figure 2. Analysis also accounts for variance in experts' responses



Analysis also accounts for variance

Note: Chart content purely notional.

Each quadrant signals the combined score of survey items along the actionability and importance dimensions. Depending on the graphical positioning, items can be considered as belonging to one of four different categories:

- **“Low-hanging fruit”**, later renamed “policy opportunities”, for items that have high scores for both metrics and could perhaps be important for action with fewer barriers to achievement then those classified as “dedicated attention needed”.
- **“Dedicated attention needed”**, renamed “heightened attention needed” for items with high importance but with a low score on perceived actionability opportunities. Thus, these items were perceived as important by the experts, but with more perceived barriers to achieving them than those discussed in the previous bullet.
- **“Nice to haves”**, for items with high actionability which are perceived as easily materialising through policy actions, but of relatively low importance.
- **“Not worth it”**, later renamed “not high priority” for items that are considered to be of marginal importance and low actionability.

However, as seen in Figure 2, even if an item is ranked in a particular quadrant, there may be significant variance underlying the response, which should be considered in interpreting the results.

As the results of the survey presented at the expert group meeting were early and changed to account for additional submissions after the meeting, the specific results are not presented in this summary. However, final results, as well as a description of each topic, will be provided in two forthcoming reports:

- *Exploring Artificial Intelligence Futures: Prospective milestones, benefits and benefits*.
- *Exploring Artificial Intelligence Futures: Prospective solutions and governance approaches*.

Expert group views on the survey and its results

In general, experts appreciated the design of the survey and the universe of options presented for consideration. However, some members critiqued the length of the survey, which they suggested may need to be reduced if it were to be leveraged for a wider audience. Further thoughts advanced by expert group members can be grouped thematically in the following points:

Clarify the meaning of low actionability when communicating results. Experts expressed their view on how to consider items with a low rating on actionability. They suggested clarifying that that low actionability does not signal lo priority of a consideration for a survey item, but only a perceived difficulty in taking action to realise it. The OECD Secretariat responded that will incorporate this feedback and clarify the meaning of actionability in forthcoming reports the leverage the survey results.

Further explore disagreement among members. Experts expressed that one of the strengths of the group is its diversity of perspectives. Hence, they encouraged exploring areas of disagreement in order to promote fruitful discussions and to gain and fuller understanding of the context of these areas. Experts suggested to consider exploring three approaches:

- **Surveying experts again at a later date** on the extent to which, if any, they have had a change of opinion about survey items over time. The OECD Secretariat noticed this comment and will consider its feasibility and usefulness for future efforts.
- **Following the Oxford Process** working method, in which topics of scientific controversy disagreement are debated until scholars agree on a description that is a fair representation of disagreement. The OECD noted this comment, and also indicated that expert group members would receive drafts of any future publications for their review to help ensure the topics and potential debates around them are described in a way that everyone can agree is fair.

- **Discussing areas of disagreement** to help identify and characterise the source of disagreement concerning specific considerations (e.g., variance in scores in some areas). The OECD Secretariat noted the importance to further investigate these areas in future meetings and scenario exploration exercises.
- **Consider the impact of AI progress' speed.** Concerning the nature of AI change, it is paramount to understand both the development speed and impact of this technology. While the survey prominently focused on impact, it did less so on speed. However, this dimension is noteworthy as it has the potential to influence the effectiveness of human responses based on the pace of development. For instance, gradual technological changes would be associated with more capacity to respond, at parity of potential predicted impact. The OECD noted that it would conduct future "super forecasting" efforts focused more on time horizons.
- **Consider the human dimension when assessing impacts.** Equally important to accurately assess impacts is the assumption on the nature of human behaviour associated with the management of the technology. In particular, at the micro level, assuming that individuals have the ability to learn and adapt would yield different impacts than a scenario in which humans are static beings that slowly react to change. At the macro-level, this consideration translates in observations about the institutional capacity of societies.
- **Foster a meta-analysis approach.** Some experts expressed that humans tend to be biased towards things they know, so it is reflected in initial results. AI "safety" , for instance, is something one can grasp, so some responses in this area appeared stronger. This may also be seen in the somewhat higher level of importance given to risks compared to benefits. This can be a limitation for future-oriented efforts, as it is challenging to get people to think beyond concepts with which they are already familiar. It was suggested that, in addition to the survey, the OECD could try to create a handful of visions of the future to help ground thinking. The OECD discussed that future efforts would narrow the scope of focus over time and orient more around hypothetical future scenarios to help ground thinking and discussion.
- **Consider "funding and adoption" as an isolated metric.** Some experts expressed that sustainable funding of solutions, which was a consideration under the "actionability" axis could be a separate item of evaluation, and that adoption (e.g., by cities, businesses and communities) is an important question that was not explicitly mentioned under the actionability item and should be discussed.
- **Increase focus on current technical trends**. One of the experts argued that the group should spend some time assessing the new stage we have entered of "artificial capable intelligence" (ACI), suggesting that new developments in AI in the present and near-term are underappreciated relative to longer-term discussions.

Topic 2: Potential futures and scenarios

OECD Secretariat presentation

Based on previous discussions and the survey results, Hamish Hobbs from the OECD Strategic Foresight Unit presented views of desirable and undesirable futures, depicting the most important and sufficiently actionable risks and benefits (Table 1). Threads were drawn from the top 20% most important items from the initial survey results that were also on the top half the actionability scale.

Table 1. Desirable and undesirable futures based on survey results

Desirable futures	Undesirable futures
Controlled training and deployment of high-risk models and applications.	**Lack of governance** of high-risk models and applications.
Strong technical tools for safe and ethical AI.	**Lack of technical tools** for safe and ethical AI.
International cooperation to ensure safe and ethical AI.	AI is frequently misused by **malicious actors.**
Widely **distributed AI benefits.**	**International competition** drives a race to the bottom on AI safety and ethics and prevents collaboration on shared rules.
Controls to **prevent excess power concentrations.**	AI benefits accrue to a **select few.**
Empowered public with strong democratic and civil society oversight.	AI **concentrates power** in harmful ways.

Mr. Hobbs explained that these envisioned futures would be used to develop scenarios for a scenario exploration exercise for the next meeting of the Expert Group on 9 November. The scenarios will be used to identify which policies are required to push us towards the beneficial futures and away from the harmful futures.

Expert Group members were invited to react to the current key threads identified by the OECD Secretariat, to better inform the scenario development.

Expert group members' views

Experts reacted to the potential futures, further analysing the consequences that certain AI applications might be associated with in a future society. They also suggested ways to develop the concepts into complete scenarios. The insights can be categorised in (I) areas of convergence or agreement, (II) areas of disagreement, and (III) other insights.

Areas of convergence of experts' opinions

Create inclusive scenarios. To better inform the general public and reach a wider audience, experts suggested to increase contextual clarity when developing scenarios. Given the differences between countries in terms of industrialisation level or regulatory culture, experts suggested to better define the contextual nuances of information asymmetry or concentration of power considerations. In this regard, it is crucial to explicitly outline the implications for society and the rule of law, when some countries are not represented by governance systems or when minorities are excluded from a certain policy intervention. These considerations must hence be included to have future scenarios that are inclusive and understandable.

Explore origins of consensus. Several experts expressed their satisfaction to see where the consensus has emerged to date. They suggested to keep this consensus still at a high-level, opening space rather than diving into the specifics of AI impacts considerations, which generally tend to be more divisive. For instance, survey's results signal particular convergence on controlling dangerous models while also mitigating concentrations of power. Experts also seemed to agree, for instance, that a full ban on advanced AI would be detrimental while more controlled release protocols might be beneficial. In this context, the need for further and constant research on high consensus items has been depicted as having the potential to bring essential insights to the group.

Areas of divergence of experts' opinions

Analogy between nuclear and AI technologies.

In the context of potential governance solutions to mitigate the negative impacts of AI, promoting international cooperation and regulation along the lines of what has been treaties as has been done with nuclear weapons is often presented as a potentially effective approach to mitigate AI risks.

- However, **one expert noted how the comparison between AI and nuclear power is potentially detrimental to effective progress**. According to the expert's view, the analogy breaks down very quickly, given the invisible nature of AI harm as opposed to nuclear weapons. In particular, as AI is woven into the fabric of society, its effects are neither physical nor perceptible, as in the case of nuclear systems. Hence, scenarios and proposed governance responses should clarify.
- Other experts, though, noted how **radioactive radiation is also invisible**, it is important to clarify the similarities and differences between nuclear and AI technology, when drawing this analogy.

Approaches to AI regulation

- **One expert suggested considering policy and governance responses whose scope is beyond strictly AI regulation.** For instance, data protection legislation may represent the most effective way to mitigate or manage some of the negative impacts of AI. Generally speaking, holistically addressing issues - for instance by promoting individual rights - is a promising approach that protects human rights including in and beyond the AI sphere.
- **However, another expert suggested focusing more on fostering governments' capacity to regulate AI specifically*.*** According to their view, for positive futures to materialise, it is fundamental that governments have a capacity to regulate. To foster this capacity, there are four conditions to be met. First of all, it is essential that resources be directed at developing technical capacity, increasing budgets, and ensuring authority. Secondly, the legal authority to impose penalties must be ensured. Thirdly, there has to be enough international collaboration to prevent substitution effects. Finally, an essential element is the existence of political support in communities around the world.

AI governance and AI safety

- According to one expert, **governance debates on broad societal impacts should be prioritised over narrower considerations on AI safety*.*** Concentrating the efforts and debates around governance keeps the conversation broader to include all options available to institutions to manage AI deployment in society. Furthermore, as the Global South has very different governance needs, these should be considered.
- On the other hand, one expert highlighted how, despite agreeing that focusing on governance should be the final goal, **incentivising AI safety represents a fundamental step to increase and foster the regulatory capacity of governments*.***

Other insights

Alternative political processes

According to one expert, in a democracy, it is essential that AI systems are developed and deployed in a manner that satisfies the majority, and not necessarily the totality, of the population.

By integrating AI in society once a general level of agreement is reached, it would be possible to create very large gains for the general population in a fast-paced manner, to then focus on how to accommodate the needs of the remaining part of the population. This would represent an antithetical approach to the inefficient one characterising present decision-making structures. The expert argued that current political processes focus on solutions that are a compromise between opposed extremes of opinion, in order to satisfy all parties in society. However, this approach results in slow implementation processes that impede society from reaping the benefits of AI in a reactive way.

Consider uncertainty and context as key dimensions to build scenarios

One of the experts identified two varying aspects that differentiate scenarios from one another. On one hand, different scenarios vary in terms of uncertainty, meaning that the risks and benefits that characterise them differ in terms of materialisation likelihood. In scenarios where certain risks have already or almost concretised, it was suggested to focus on the evolution of the said risk.
On the other hand, scenarios should take into account dimensions of context, as countries around the world greatly differ in terms of industrialisation level or political ideology. Hence, the same scenario would evolve or materialise in different ways based on the political, geographical and socio-economic context that is assumed.

Topic 3: Workshop on AI in Work, Innovation, Productivity & Skills (AI-WIPS)

OECD Secretariat Presentation

Alistair Nolan from the Science and Technology Policy (STP) division of the Science, Technology and Innovation Directorate (STI) presented on the potential for a joint workshop (tentatively anticipated to occur in Q2 2024) with the OECD.AI Policy Observatory and OECD Strategic Foresight Unit that could involve members of the expert group. He presented on a handful of potential workshop topics to obtain expert opinions on the perceived relevance of the topics, and to gauge members' initial interest in participating. It was explained that additional topics may be added to the workshop depending on expert group progress and interests at the time it is organised.

In particular, his presentation focused on four macro-themes: AI for better policymaking; AI in materials sciences; AI and socially productive robots; AI and possible dual use in science.

AI for better policymaking

In the context of policymaking, large language models (LLMs) and other generative AI models could play a crucial role in informing policymakers. For instance, they could be used to produce accurate forecasts in real-time, simulate scenarios, as well as use diverse sources of qualitative data to produce output. Furthermore, Mr. Nolan highlighted how collective intelligence integrated with AI systems is a fruitful investment area which can drive fast progress.

AI in materials sciences

Machine learning has already been used for a while in materials sciences, a scientific field focussing on researching and discovering novel materials to drive innovation.
Challenges in the field exist, with protein folding types of problems being a primary example, where AI systems are not yet fully able to carry out complex calculations. Breakthroughs in the field have a great deal of potential, for instance, by aiding climate change efforts by optimising manufacturing of ultra-light materials.
Further beneficial impacts of AI could be experienced in the geopolitical domain. As the domestic semi-conductor manufacturing industries in the US and EU heavily rely on China's economy of scale concerning critical minerals extraction, AI could develop alternative processes to make smaller scale extraction

profitable. Similarly, its use might aid development efforts towards digital twins of material processes that do not yet exist in order to discover alternative and more affordable materials to be used in place of critical raw materials.

AI and socially productive robots

The case of socially productive robots should be analysed under the COVID-19 context. During the crisis, socially productive robots were considered as a highly effective tool to be employed in hospitals. However, several cost-benefit analyses performed to assess whether to deploy robotics were mostly negative. An example case is represented by disinfection robots that were employed during the crisis, which presented substantial difficulties in navigating the physical environment and processing COVID-19 test samples because of their different shapes. Generally speaking, robots cannot yet deal with the social environment due to a large amount of variables, and this seems unlikely to change soon.

Dual use of AI

Mr. Nolan finally presented the risks of dual use in AI. For instance, AI models can be used in chemistry labs to minimise toxicity of materials; however, one could also maximise the toxicity of materials, by simply flipping algorithm and with the aid of publicly available data.

Experts feedback and conclusion

Experts subsequently reacted to the content of the presentation and shared the following insights:

- **Investment in AI forecasting.** In the context of AI in policymaking, governments should start investing in AI technology as it represents a fruitful area for public sector's service improvement in terms of governance. Furthermore, the work of the group has been highlighted as a unique instrument to stimulate change, in areas where the market or states alone will not be able to achieve the broad impact they wish achieve.
- **Focus on future-oriented work on neuro-technologies.** AI combined with other technologies, such as neurotech, has high potential to create societal benefits and should hence be the focus of OECD work.
- **Biomedicine norms as solid basis for AI work.** Bioengineering and biomedicine community have developed a set of norms that are powerful enough to be leveraged in the work of the Expert Group.
- **Asymmetric benefits in LLMs.** LLMs performances and outcomes are heavily shaped by data used for training. When English language is not the base, a form of dual use could materialize, in the sense that ASEAN countries need to develop different LLMs to reap the benefits, or they will incur in negative impacts.

Annex A: Elements and early findings of the expert group survey

The following items were presented to expert group members based on early survey results. Additional responses were received after the expert group meeting. As a result, some of the categories for some items may have changed. The final responses will be published in two forthcoming OECD reports, tentatively titled:

- Exploring Artificial Intelligence Futures: Prospective milestones, risks and benefits.
- Exploring Artificial Intelligence Futures: Prospective solutions and governance approaches.

Annex A - List of Participants

Rebecca Anselmetti - Senior Policy Advisor at the UK government's Department for Digital, Culture, Media and Sport.

Azeem Azhar – Founder of Exponential View.

Joscha Bach - Principal AI Engineer at Intel Labs.

Amir Banifatemi – Co-founder, AI Commons.

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

Jamie Berryhill – AI Policy Analyst at OECD.

Nozha Boujemaa - Digital Trust Officer at Decathlon.

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

Rumman Chowdhury - Responsible AI Fellow at Harvard University's Berkman Klein Center.

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

Mariano-Florentino (Tino) Cuéllar - President and CEO at Carnegie Endowment for International Peace.

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.

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

Juha Heikkilä - Head of Unit, Robotics, Directorate-General for Communication Networks, Content and Technology at DG CONNECT, European Commission.

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

Emmanuel Kahembwe - CEO - VDE UK.

Holden Karnofsky - Director of AI Strategy at Open Philanthropy.

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

Rafal Kierzenkowski - Senior Counsellor for Strategic Foresight at OECD.

Daniel Leufer - Senior Policy Analyst at Access Now.

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 the OECD.AI, OECD Digital Economy Policy Division – OECD.

Andrea Renda - Senior Research Fellow and Head of Global Governance, Regulation, Innovation and the Digital Economy (GRID) - Centre for European Policy Studies.

Francesca Rossi - IBM Fellow and AI Ethics Global Leader, IBM.

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).

Toby Walsh - Chief Scientist of UNSW.AI.