Expert Roundtable on Generative AI

Context

In just a few months, generative AI has gone from technical lab discussions to daily front-page news. It is already used in many sectors to create individualised and scalable content, automate tasks, and improve productivity, and it has the **potential to revolutionise industries and society**.

However, the full impact of generative AI, including **potential risks and ramifications, is not fully known**. One of the biggest risks stems from its capacity to generate manipulated content such as mis- and dis-information and deepfakes. These could provoke **severe negative consequences,** such as serious social, political, and economic repercussions at scale.

Recognising the transformative and disruptive potential of generative AI, the G7 encouraged the Organisation for Economic Co-operation and Development (OECD) and other international organisations to promote international co-operation and explore relevant policy developments and practical projects, including disinformation issues. The G7 also asked the OECD to contribute to a new Hiroshima AI Process of international discussions on AI.

Pursuant to this mandate, on 31 May 2023, the OECD.AI Policy Observatory and Strategic Foresight Unit held a virtual Expert Roundtable on Generative AI. The roundtable comprised the new OECD Expert Group on AI Futures, including its co-chairs, members of the secretariat, and membership candidates. The co-chairs are:

- 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 Strategic Foresight and Analysis, German Federal Ministry of Labour and Social Affairs.

The discussion focused on three key questions:

1. Given ongoing discussions, how might governments consider rules for the **controlled release of high-impact generative AI models and applications**? To guide this, how can we better measure/assess the quality and accuracy of generative AI outputs?
2. How can we promote **regulatory experimentation** that benefits countries and companies?
3. How can we better **measure/assess the capabilities of generative AI relative to humans?** How do we measure/assess how these capabilities could impact societies and labour markets?

Along with the co-chairs, 14 experts participated in the discussion and expressed their views on each topic, under the Chatham House Rule. Although this summary is non-attributional in accordance with the Chatham House Rule, the participants agreed to having their names published in the List of Participants below.

In addition to the room discussion, participants were able to submit brief documents to outline their views on each topic. Contributions from these documents are woven into the summary below.

Controlled release of high-impact generative AI models and applications

Speakers highlighted several points and fundamental approaches to mitigate risks when releasing foundation models or other potentially high-impact generative models:

- **Preview use, testing, and transparency** are fundamental features of a regulatory approach to oversight before deployment. Certain stakeholders would get some degree of access to evaluate the model and flag potential shortcomings. Those stakeholders could include regulators, researchers, advocacy groups and auditors.
- **Evidence standards** are used in other domains and could be a useful pre-release strategy. For example, in healthcare, standards specify what evidence digital health technology developers can use to demonstrate benefits. Evidence standards specify what developers need to show to justify effectiveness and how they mitigate harm. This could be tailored for particular domains or more broadly for general-purpose models. Requiring developers to demonstrate evidence shifts more of the burden onto them and from external stakeholders, such as academic or political institutions. Oversight methods only allow API access and not full model access, collaborative work with the developers, or transparency over data. This restricts the possibility of independent evaluation.
- **Regulation should focus on undesirable usages that fall outside the functional capabilities of the system.** Types of usage that fall into this category include using ChatGPT to answer a factual question when the system is, in fact, unable to complete the task reliably. In this context, honest advertising should play a prominent role in mitigating disinformation risks and should thus be fostered. Similarly, there must be programs to promote users’ education and systems that allow them to flag illegal or undesirable content.
- **Controlled model releases can be paired with rules for applications built using those models.** Powerful models are already in circulation through a mix of deliberate releases and open-source efforts, and even commercial leaks. Release rules for models can be stronger if corresponding rules are applied downstream to applications. For instance, governments could apply transparency and fairness requirements for underlying data sets to the quality of the applications’ output, propagating the same requirements up the value chain.
- **Participants identified standards as a crucial tool for effective regulation.** Interestingly, participants noted that standards rarely reflect global values. Nor are they a universal solution to multiple problems, as challenges might differ depending on the contexts they characterise. Hence, it is important to rely on flexible approaches that adopt and promote different requirements for each ecosystem level, e.g. data, cloud provider, etc. The focus, in particular, should be directed at fostering interoperability, consensus and consistency at the international level.

How to measure the accuracy of generative AI outcomes

During the discussion on the controlled release of high-impact generative AI models, participants sometimes expressed differing views on the importance of model “accuracy”. In general, speakers agreed on the role of accuracy as a critical evaluative factor, while they also acknowledged it is rarely the only important performance assessment metric. The main discussion points were:

- **Accuracy is rarely the only desired goal.** When deciding on a model or application release, understanding potential *harms* requires much more than accuracy. Those reviewing models should consider additional broad categories of impact.

- **In the context of generative AI, extrapolation mechanisms based on probabilistic inference by definition cannot always be accurate.** Instead of focusing on accuracy metrics, we should be evaluating to determine the system's quality of output. We should also explore which methods may be most effective in mitigating any negative and harmful consequences, such as the generation and diffusion of hate speech. During testing, market actors should not be allowed to test their own systems in an isolated and independent manner. Rather, experts should design proper, independent testing and guidelines for these systems.
- **Quality and accuracy requirements need to be based on metrics and test tools that need to be developed urgently,** e.g. in a combination of academic research and standardisation. The OECD is developing its own too, the OECD Responsible Business Conduct (RBC) guidelines on AI risk. Once they are released, anyone developing metrics and tools should refer to them. The scope of some of these quality metrics will go beyond generative AI models and cover the combination of the model's characteristics plus the competencies and role of the human user/operator.

Some novel considerations underlined the role of cultural and sociotechnical perspectives when assessing accuracy. In particular, some participants stressed that:

- In the context of cultural representation and accuracy, **foundation language models display a lower level of accuracy with respect to non-English languages**. Furthermore, regulation in the field is usually associated with data governance and protection laws, which might prove problematic from a global perspective. Many countries do not possess the technical means or know-how to impose and enforce these rules to increase the model's accuracy and degree of representativeness.
- **Human evaluation and sociotechnical analysis in context are crucial**. Referring to definitions of *intrinsic evaluation*, e.g. benchmarks, and *extrinsic evaluation*, i.e. research in the field to understand the whole system in its particular setting, we need both. Intrinsic evaluation is critical to chart progress, and extrinsic evaluation to understand real-world performance and impacts better. Ideally, we should design intrinsic evaluations that have some signal to the extrinsic evaluation and iterate over time to ensure they stay aligned.

Words of caution

As part of the discussion on the controlled release of high-impact generative models, participants expressed words of caution, either about things to consider when exploring the controlled release of high-impact generative AI systems or about harms that this approach may not be able to mitigate. In particular:

- **Malicious uses of the technology should be regulated**. The main difficulty with generative AI creating harmful images, texts, or videos is the very low cost and velocity with which they can be created. This means it may be necessary to punish the concerted creation of disinformation through generative AI more severely than when created by traditional means.
- **The use of generative AI may need to be treated differently in the context of critical infrastructure and high-impact areas such as healthcare.** Participants emphasised that regulating the technology's application and specific use cases rather than the technology itself would avoid negative consequences in the short term.
- **Participants cited a lack of effective enforcement mechanisms as the main challenge when dealing with AI system regulation.** it will be critical to think of ways to effectively enforce rules, whether related to controlled release or beyond that, as the field has more governance obstacles than technical ones.

Main points from written contributions

Participants were invited to submit written briefs to outline and further explain their points, or to raise additional issues they may not have had an opportunity raise during the discussion session (e.g., due to time constraints). Key points were:

- **Avoid large, bulky regulatory regimes, particularly horizontal legislative approaches that could stifle innovation**. Instead, regulators could prioritise vertical approaches for generative AI models and applications. The benefits of a vertical approach are that it offers greater regulatory precision and can pave the way for specific remedies to particular problems. The drawback is its piecemeal nature, which forces regulators to devise new rules for new applications or problem sets as they emerge.
- **Focus on rules for applications built on top of generative AI models.** Governments could, for instance, enforce requirements on transparency and fairness of underlying data sets orthe output quality. As a result, the requirements would propagate back up the value chain effectively and in a desirable way.

Regulatory experimentation that benefits both countries and companies

The second topic of the session was the importance of fostering regulatory experimentation that maintains a balance between private incentives and public interests. Participants shared what they believe would be effective:

- **Promote diversity**. The impacts of tech are context-dependent, so it will be important to think about geographical, cultural and economic distinctions. Any international agreement should not constrain suitable development choices for other parts of the world. Promoting successful regulatory experimentation may require anapproach like cybersecurity that identifies emerging risks and flows into decision-making processes.
- **Innovation should move beyond fines as the default stick in regulation.** As big corporations have significantly higher profits, fines alone are not proving to be effective in achieving regulators' goals. Legislators should aim to prevent companies from committing illegal practices by developing effective measures such as algorithmic disgorgement. This measure, also called *algorithmic destruction* and *machine learning model deletion*, is an enforcement tool requiring organisations to delete machine learning models and algorithms developed with flawed data.

Other important insights concerned novel approaches to regulatory experimentation:

- **Measuring public and individual opinions over time and on a continuous basis** is crucial in a society where people are interacting with AI-enabled chatbots, and the associated psychological and political effects can be negative. Because chatbots have opaque objective functions and are trained on a vast quantity of data, they can have very unpredictable results. Therefore, governments need to think of real enforcement mechanisms to limit the spread of misinformation and malicious or illegal use of generative technologies.
- **Developing quality standards for regulatory experimentation.** This fundamental aspect is crucial but often ignored in assessing the effectiveness of regulatory experimentation. In the field of pharmaceuticals, there are global and clear standards to ensure quality in experiments.

Governments should strive to reach a similar methodology for generative AI, whereby comparative elements are available to assess the effectiveness and validity of experimental results. The proposed example concerned the application of Randomised Control Trials (RCTs) to AI applications in large corporations as a successful approach. Embedding this approach in the policy learning process could be useful and done through regulatory sandboxes. Participants also said the cross-border nature of technological innovations should skew governmental efforts towards increased regulatory cooperation and experimentation with open standards.

- **Regulatory experimentation should focus on infrastructure and technical protocols used in the digital domain.** It is essential to develop effective technical measures which make it unfeasible to develop and deploy certain uses of generative AI instead of prohibiting them.

Main points from written contributions

- **Prioritise innovation flexibility**. There is understandable concern about the potential risks of generative AI products and the societal harms they may cause. At the same time, regulators should be careful not to let AI "panic" supplant an open and flexible regulatory regime. Regulators should investigate new techniques for enhancing AI safety and encouraging responsible use. They could consider iterative oversight approaches that allow for more immediate testing and responses to new innovations or products that hit the market. Rather than devise blanket rules or force products to undergo lengthy regulator certifications, policy makers should establish structures that allow for sufficient experimentation and guard against major risks. One way to achieve this balance is to establish ex-ante, i.e. before the event, risk assessment models. For applications with higher potential risks, regulators could mandate certain restrictions, such as extended testing periods or establishing regulatory sandboxes.
- **Design regulatory systems to prevent substantial harms**. Not all AI systems bring equivalent risks. Certain general-purpose models may incorporate dangerous capabilities inadvertently or by default. Regulators should incorporate special processes for such products. First, regulators could mandate developers to evaluate AI systems for "dangerous capabilities and alignment." This will force software engineers to consider and identify risks early and allow them to become more responsive to threats when training new models or deploying applications. Second, regulators can move beyond a voluntary compliance regime. The threat of harm is too great to trust that developers will properly incorporate risk assessments before bringing products to market. Regulators could be much more hands-on throughout a product's life cycle. They could mandate, for example, that companies file mandatory risk evaluations and establish special procedures for applications deemed sufficiently risky, e.g. requiring enhanced testing or oversight.

How to measure the relative capabilities of generative AI compared to humans

As a note, the OECD released its Employment Outlook 2023 in the weeks following the roundtable discussion. It focuses on the impact of AI on jobs and touches on issues relevant to this discussion. At the roundtable itself, participants provided a number of thoughts on measuring machine capabilities, often turning directly to how increasing machine capabilities may impact labour markets.

- Speakers highlighted several possible approaches to measure the relative capabilities of generative AI. Some common insights that speakers seemed to agree on concerned the role of big job platforms as sources of preliminary data and the key role of costs, respectively. **Big job platforms such as LinkedIn, Glassdoor, and Indeed can be used as data sources to better understand the relative capabilities of generative AI compared to humans**. In particular, by analysing the types of tasks demanded in different job postings over time, it would be possible to estimate intra-sectoral changes related to human and generative AI capabilities relative to each other.

- **Job platforms were also identified as a meaningful source for labour impact analysis**, specifically as they provide timely and accurate wage data across different sectors. Monitoring wage differentials over time might reveal an effective way to detect early signs of downward pressure on wages, as this has historically been the effect of technologies deployed in the market. Similarly, analysing the frequency of job openings or the evolution in types of job offers on big platforms could be insightful in assessing the impact of AI on labour market outcomes.
- **Costs are paramount in determining the technology's degree of deployment and associated labour market impacts.** The potential impact of generative AI on the labour market is not only driven by AI tech capabilities but also by costs. The first point raised by participants concerned outsourcing options. For many years, it has been common to outsource production to countries with low labour costs, which comes with the difficulty of quality control and cultural and jurisdictional clashes. Generative AI capabilities may change this by replacing some outsourced tasks that do not demand high quality. The second point concerned issues posed by the underlying cost structure. Developing generative AI systems is costly and characterised by extreme resource intensiveness. However, costs are decreasing, which may result in more companies leveraging the technology. The results of this on the labour market remain to be seen.

Other important concepts outlined during the session concerned the role of benchmarking, labelling requirements, and public engagement in the definition and analysis of AI capabilities.

- While AI and human capabilities can be compared via benchmarking, the larger societal implications must also be considered when defining benchmarks. AI/Human trade-offs may have short- and long-term impacts like job displacement, skills gaps, income inequality, and privacy issues, among others. AI safety research plays a crucial role in identifying potential risks and developing mitigation strategies, and it should be further incentivised to include broader societal and economic implications.
- Labelling has been defined as one key type of regulation to define AI capabilities. Specifically, one suggestion proposed systems should identify themselves as machines when interacting with humans and use labels to identify AI-generated content. For instance, proposals exist to create repositories of encrypted data that come from generative engines to trace the content’s origin.
- On a contrasting note, a more sceptical view highlighted the shortcomings of labelling protocols, especially under the circumstances where the tasks’ degree of automation is not clearly defined. In the case of customer support services, for instance, there is a full spectrum of automation degrees in the process. As human operators can either fully rely on AI or be partly assisted by automated systems, it would be virtually impossible to effectively label tasks which tasks are performed by AI. In this context, the question of responsibility and interaction between humans and technology is a central one, which should be addressed through clearly developed guidelines.
- Citizens’ and workers’ input should shape policymaking for benchmarking and labelling. It is important to map which approaches and tools can be applied within companies to develop best practices and accountability in the dynamics between AI and humans in the labour market. The end goal would be ensuring that

citizens’ and workers’ perspectives are factored in into the policymaking process, fostering the emergence of a standardised accountability process. This will require extensive, open public and labour worker engagement.

Main points from written contributions

Key points from written submitted briefs on generative AI’s impact on labour markets:

- Critically, **society needs to account for how jobs may be redefined to effectively deploy the technology** and avoid scenarios that lead to widespread poor working conditions. While certain jobs may be very hard to automate, some companies have redesigned the pipeline to make tasks easier to automate. For example, it's difficult to automate a shop assistant, so instead, Amazon created a system in which many parts can be automated, and the remaining tasks can be algorithmically managed. Such changes have often been associated with degrading pay, working practices, and worker rights.

List of Participants

Co-chairs

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, German Federal Chancellery.

Participants

Miguel Amaral, Senior Economist, Public Governance Directorate, OECD.

Carolyn Ashurst, Senior Research Associate in Safe and Ethical AI at the Alan Turing Institute.

Azeem Azhar, Founder of Exponential View.

Amir Banifatemi, Co-founder of AI Commons.

Farzana Dudhwala, AI Policy & Governance Manager at Meta.

Steven Feldstein, Senior Fellow at Carnegie Endowment for International Peace (written contributions).

Marko Grobelnik, AI Researcher at Artificial Intelligence Lab at Jozef Stefan Institute.

Sebastian Hallensleben, Chair of CEN-CENELEC JTC AI at the VDE Association for Electrical, Electronic & Information Technologies.

Stephanie Ifayemi, Head of Policy of Partnership in AI.

Daniel Leufer, Senior Policy Analyst at Access Now.

Vukosi Marivate, Chair of Data Science at University of Pretoria.

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

Roman Yampolskiy, Computer Scientist at University of Louisville.

Katharina Zweig, Professor and Researcher at TU Kaiserslautern.