OECD

OECD.AI Policy Observatory

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 has the **potential to revolutionise industries and society** and is already being used in a variety of sectors to create individualised and scalable content, automate tasks, and improve productivity.

However, the full impacts and **potential ramifications of generative AI are not fully known**, and the technology be misused with **severe negative consequences** through mis- and dis-information, deepfakes, and other manipulated content. This can lead to serious social, political, and economic repercussions at scale.

Recognising the transformative and disruptive potential of generative AI, the G7 charged the Organisation for Economic Co-operation and Development (OECD) and other international organisations with promoting international co-operation and exploring relevant policy developments and practical projects, including on issues of disinformation. The G7 also charged the OECD with contributing 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 leveraged the co-chairs, infrastructure, and membership candidates of the new OECD Expert Group on AI Futures. that were in turn presented by the three co-chairs of the Expert Group on AI Futures, who are namely:

- 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 relevant to generative AI:

1. Given the on-going discussion whether governments should consider rules around **controlled release of high-impact generative AI models and applications**, how might this be done? 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 both countries and companies?
3. How can we better **measure/assess the relative capabilities of generative AI compared to humans** and how these capabilities may impact societies and labour markets?

In addition to the co-chairs, 14 experts participated in the discussion and expressed their on each respective topic under the Chatham House Rule. Although this summary is non-attributional in accordance with the Chatham House Rule, the participants agreed that their names be published, which can be found in the List of Participants section below.

In addition to the discussion in the room, participants were also afforded the opportunity to submit brief documents outline their views on these topics. 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 that could be used to mitigate risks when releasing foundation models or other potentially high-impact generative models. In particular, participants raised the following views:

- **Preview use, testing, and transparency** are fundamental features of a regulatory approach which aims to promote oversight before deployment, whereby certain stakeholders are given some degree of access to evaluate the model and flag potential shortcomings. Stakeholders could include regulators, researchers, advocacy groups and auditors.
- **Evidence standards** are used in other domains: (e.g. in the healthcare sector, where standards specify what evidence developers of digital health technologies can use to demonstrate benefits) and could be a useful pre-release strategy. Evidence standards specify what evidence developers need to show to justify effectiveness and mitigations of harm. This could be tailored for use in particular domains, or more broadly for general purpose models. Requiring developers to demonstrate evidence shifts more of the burden onto developers, rather than external stakeholders such as academic or political institutions. In this context, current oversight methods allowing for API access alone, but neglecting full model access, collaborative work with the developers, or transparency over data, restrict the ability for independent evaluation.
- **The focus of regulation should be on the types of usage that are not desired and that fall outside the functional capabilities of the system.** Types of usage that fall in this category include using ChatGPT to answer a factual question when the system is in fact unable to reliably complete the task. In this context, honest advertising should play a prominent role in mitigating disinformation risks and should thus be fostered. Similarly, it is necessary to promote users' education, as well as to create systems that allow users to flag illegal or undesirable content.
- **Controlled release of models can be paired with rules around applications built on top of these models.** Powerful models are already out on the market and in society, through a mix of deliberate releases, commercial leaks, and open source efforts. Hence, the impact of release rules can be strengthened through corresponding rules on downstream application deployment. Governments could, for instance, enforce requirements on transparency or fairness of underlying data sets or the quality of the applications' output, which would mean that these requirements propagate back up the value chain.
- **Standards development was identified as a significant factor in the definition of effective regulation.** Interestingly, it has been noted how standards rarely reflect global values or a comprehensive universal solution to multiple problems that might differ depending on the context 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 as well as consensus and consistency at the international level.

How to measure accuracy of outcomes?

As a sub-discussion on the controlled release of high-impact generative AI models, participants expressed sometimes differing views on the importance of model "accuracy". In general, speakers agreed on the role of accuracy in being a critical evaluative factor, while they also highlighted how this factor is rarely the only important performance assessment metric. The main discussion points were:

- **Accuracy is rarely the only desired goal.** When deciding on model or application release, an understanding of potential *harms* goes well beyond accuracy and should consider more broad categories of impacts.

- **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 for the system's quality of output and explore which methods may be most effective in mitigating negative and harmful consequences that might arise, such as the generation and diffusion of hate speech. In the context of testing, actors in the market should not be allowed to test their own systems in an isolated and independent manner. Rather, proper, independent testing and guidelines should be designed for these systems.
- **Quality and accuracy requirements will need to be based on metrics and test tools that don't exist yet but need to be developed urgently** (e.g. in a combination of academic research and standardisation). It may be desirable to refer to them the future OECD Responsible Business Conduct (RBC) guidelines on AI risk. The scope of some of these quality metrics will not be the generative AI model alone but the combination of characteristics of the model plus the competencies and role of the human user/operator.

Some novel considerations on this topic 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 data protection laws, which might prove problematic from a global perspective, as many countries do not possess the necessary means (both technical and related to know-how) to either impose or enforce these types of rules aimed at increasing the model's accuracy and degree of representativeness.
- **There is a strong need for human evaluation, and sociotechnical analysis in context**. 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 needed to chart progress, and extrinsic evaluation is needed to better understand real-world performance and impacts. Ideally, we should design intrinsic evaluations that have some signal to the extrinsic evaluation, and iterate over time to ensure they stay aligned.

Cautions from the participants

As part of the discussion on controlled release of high-impact generative models, an additional sub-discussion came to the fore in which participants provided cautions related to generative AI that either need to be considered when exploring controlled release of high-impact generative AI systems, or harms that such an approach may still not be able to mitigate. In particular:

- **Malicious uses of the technology should be regulated**. The main problem of creating harmful images, texts, or videos by hand or by generative AI systems is the very low cost and velocity with which they can be created. Thus, it may be necessary to punish the concerted creation of disinformation through generative AI in a manner harsher than traditional creation of such content.
- **The use of generative AI may need to be treated differently in the context of critical infrastructure or in high-impact areas such as healthcare.** Participants emphasised a need to regulate the application and use-cases of the technology rather than the technology itself, in order to concretely avoid negative consequences in the short-term.
- **The lack of existing effective enforcement mechanisms has been noted as the main impeding factor when dealing with regulation of AI systems.** Including and beyond any rules related to controlled release, it will be critical to reflect on potential ways to effectively enforce rules, as the field is characterised by the presence of governance obstacles rather than technical ones.

Live discussion supplemented with written contributions

In addition to the main points raised during the roundtable discussion, participants were invited to submit written briefs to outline or 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 from these consist of:

- **Avoiding large, bulky regulatory regimes, particularly horizontal legislative approaches to avoid stifling innovation**. Instead, regulators could consider prioritizing vertical approaches to regulating generative AI models and applications. The benefits of a vertical approach is that it offers greater regulatory precision and can pave the way for specific remedies geared towards particular problems. Its drawback is its piecemeal nature, which forces regulators to devise new rules for new applications or problem sets as they emerge.
- **Focusing on rules around applications built on top of generative AI models.** Governments could for instance enforce requirements on transparency or fairness of underlying data sets or the quality of the output. As a result, the requirements would propagate back up the value chain, in an effective and desirable manner.

Regulatory experimentation that benefits both countries and companies

The second topic of the session was the importance of fostering regulatory experimentation aimed at maintaining a balance between private incentives and public interest. In this context, discussion participants advanced an array of perspectives or courses of actions they believe would be effective:

- **Promoting diversity**. Impacts of tech are context-dependent so it will be important to think about geographical, cultural and economic distinctions. Whatever international agreement may be collectively achieved and endorsed, it should not constrain suitable development choices of other parts of the world. If society is to promote successful regulatory experimentation, an approach like in the realm of cybersecurity may be needed, one 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 the regulators' goals. In this context, legislators should aim to prevent companies from committing illegal practices by developing effective measures such as algorithmic disgorgement. This measure, also known as algorithmic destruction or machine learning model deletion, has so far been used as 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 which should not be overlooked:

- **Measuring on continuing basis what is happening with regard to public opinion and individuals over time** is crucial in a society where people are interacting with AI-enabled chatbots and the associated psychological and political effects can be negative. As chatbots have opaque objective functions, and since they're trained on a vast quantity of data, they can have very unpredictable results. Therefore, governments need to think of real enforcement mechanisms, aimed at limiting the spread of misinformation as well as the malicious or illegal use of generative technologies.
- **Developing quality standards for regulatory experimentation.** This is a fundamental aspect which is often ignored but is crucial in assessing the effectiveness of regulatory experimentation. In the field of pharmaceuticals, there's global and clear standards to ensure quality in experiments.

Governments should strive to reach a similar methodology, whereby comparative elements are available to assess the effectiveness and validity of experimental results. In this context, the example that was put forward concerned the application of randomised control trials (RCTs) in AI applications in large corporations, as a successful approach. Embedding this approach in the policy learning process would reveal useful and can be done through the promotion of regulatory sandboxes. Furthermore, it was highlighted how 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.** In this context it is essential to develop effective technical measures which instead of prohibiting certain uses of the technology, make it unfeasible to develop and deploy them.

Live discussion supplemented with written contributions

Key points from written submitted briefs from the participants consist of:

- **Prioritizing innovation flexibility**. There is understandable concern about the potential for generative AI products to foster significant risk and produce societal harms. 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, for example, 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 go through lengthy regulator certifications, policymakers should establish structures that allow for sufficient experimentation while guarding 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 periods of testing or establishing regulatory sandboxes.
- **Designing regulatory systems to prioritise preventing substantial harms**. Not all AI systems bring equivalent risks. Certain general-purpose models may incorporate dangerous capabilities inadvertently or by default. Regulators should build in special processes for such products. First, regulators could mandate that developers evaluate AI systems for “dangerous capabilities and alignment.” This will force software engineers to consider and identify risks at an early stage 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 all facets of the product 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 Employment Outlook 2023 was released in the weeks following the roundtable discussion. It focuses specifically 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 particularly agree on concerned the role of big job platforms as sources of preliminary data and the key role of costs, respectively. **Big job platforms (e.g. LinkedIn, Glassdoor, Indeed) can be used as data sources to better understand the relative capabilities of generative AI compared to humans**. In particular, analysing the types of tasks demanded in different job postings over time, it would be possible to estimate within-sector 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 data on wages across different sectors. Hence, monitoring wages 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 reveal to be insightful in assessing the impact of AI on labour market outcomes.
- **The role of costs is paramount in determining the degree of the technology's 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 has come with it associated 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 with generative AI. The second point presented on the issue concerned issues posed by the underlying cost structure. Developing generative AI systems is costs and characterised by an extreme resource intensiveness. However, costs are decreasing, which may result in more companies leveraging the technology. The results of this on the labour market is to be seen.

Other important concepts have been 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 need also to be considered in defining benchmarks. The 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 related to AI, and it should be further incentivised to include broader societal and economic implications.
- Labelling has been defined as one key type of regulation in the context of defining AI capabilities. Specifically, it was suggested that systems should identify themselves as machines when interacting with humans and that labelling of AI-generated content should be promoted. For instance, proposals exist concerning the creation of repositories of encrypted data from generative engines, which can be accessed 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, it was highlighted how there exists 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 as being performed by AI or not. In this context, the question of responsibility and interaction between humans and technology is a central one, which should be addressed through a clear development of guidelines.
- Citizens’ and workers’ input should shape policymaking in this realm. It is important to map the approaches and tools applicable 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 standardized accountability process. Extensive, open public and labour worker engagement are needed to achieve this.

Live discussion supplemented with written contributions

Key points from written submitted briefs in the context of labour market impacts of generative AI consist of:

- Critically, **society needs to account for how jobs may be redefined in order to effectively deploy the technology** and avoid scenarios in which it leads to widespread poor working conditions. Recently, while certain jobs may be very hard to automate, some companies have redesigned the pipeline in ways that make tasks easier to automate (e.g. it's difficult to automate a shop assistant, but Amazon instead created a system in which many parts can be automated, and in which the remaining tasks can be algorithmically managed). Such changes have often been associated with a degradation of 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.