Competition in generative AI (European Commission)

Open Data Institute response

March 2024

About the ODI

The Open Data Institute (ODI) is an independent, non-partisan, not-for-profit organisation founded by Sir Nigel Shadbolt and Sir Tim Berners-Lee in 2012. We have a mixed funding model and have received funding from multiple commercial organisations, philanthropic organisations, governments and intergovernmental organisations to carry out our work since 2012.

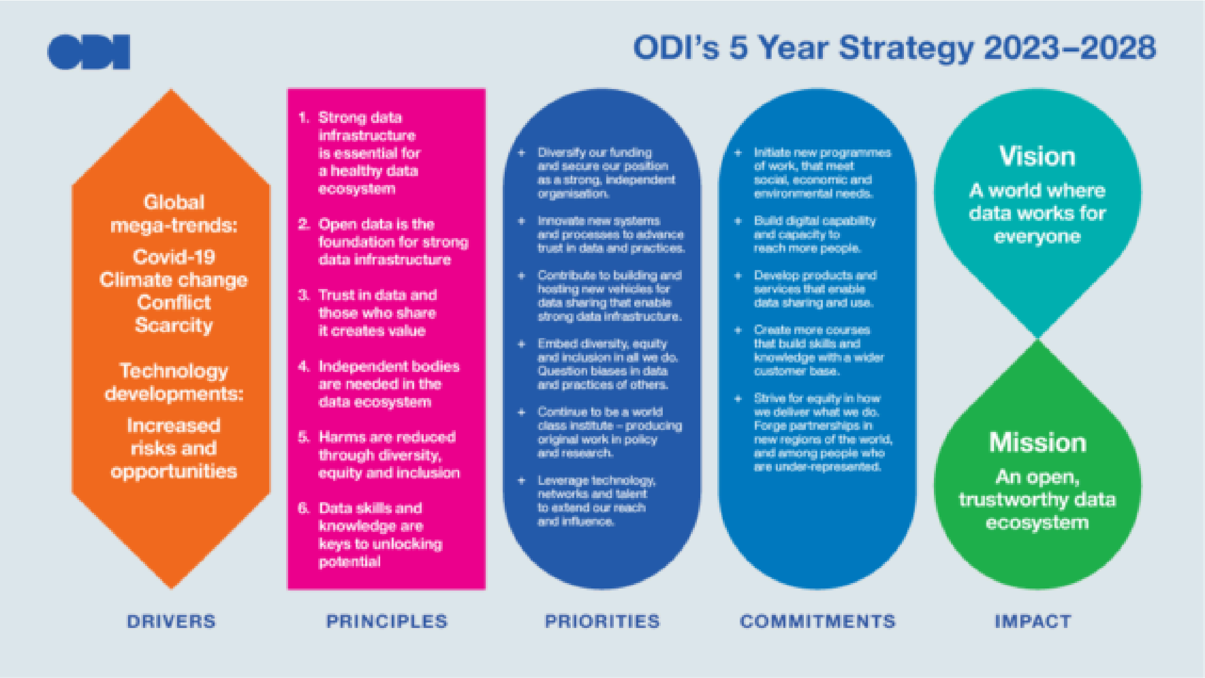
The ODI wants data to work for everyone: for people, organisations and communities to use data to make better decisions and be protected from any harmful impacts. We work with companies and governments to build an open, trustworthy data ecosystem. Our work includes:

- **consultancy:** working with organisations in the public, private and third sectors, building capacity, supporting innovation and providing advice
- **research and development:** identifying good practices, building the evidence base and creating tools, products and guidance to support change
- **policy and advocacy:** supporting policymakers to create an environment that supports an open, trustworthy data ecosystem

Our 5 year strategy sets out what we think are the elements of an open and trustworthy data ecosystem for a world where data works for everyone. Our approach allows us to adjust our implementation and engagement as the world around us, and the organisations we work with, change. Our activities will be set out on an annual basis, mapped to the six principles that guide everything we do:

1. We believe that a **strong data infrastructure** is the foundation for building an open, trustworthy data ecosystem on a global scale and that this can help address our most pressing challenges.
2. Strong data infrastructure includes data across the spectrum, from open to shared to closed. But the best possible foundation is **open data, supported and sustained as data infrastructure**. Only with this foundation will people, businesses and governments be able to realise the potential of data infrastructure across society and the economy.
3. For data to work for everyone, it needs to work across borders – geographic, organisational, economic, cultural and political. For this to happen ethically and sustainably, there needs to be trust – **trust in data and trust in those who share it**.

4. There is greater need than ever for trusted, **independent organisations to help people across all sectors, economies and societies to benefit** from better data infrastructure.
5. For data to work for everyone, those collecting and using it need to be highly alert to inequalities, biases and power asymmetries. All organisations working in data must take proactive steps to ensure that they **contribute fully and consciously to creating a diverse, equitable and inclusive data ecosystem**.
6. The world needs a new cohort of **data leaders – individuals who have data knowledge and skills** and are equipped to understand the value, limitations and opportunities offered by data, data practices and data sharing.



ODI's 5 Year Strategy 2023–2028

Global mega-trends:
Covid-19
Climate change
Conflict
Scarcity

Technology developments:
Increased risks and opportunities

DRIVERS

1. Strong data infrastructure is essential for a healthy data ecosystem
2. Open data is the foundation for strong data infrastructure
3. Trust in data and those who share it creates value
4. Independent bodies are needed in the data ecosystem
5. Harms are reduced through diversity, equity and inclusion
6. Data skills and knowledge are keys to unlocking potential

PRINCIPLES

- Diversify our funding and secure our position as a strong, independent organisation.
- Innovate new systems and processes to advance trust in data and practices.
- Contribute to building and hosting new vehicles for data sharing that enable strong data infrastructure.
- Embed diversity, equity and inclusion in all we do. Question biases in data and practices of others.
- Continue to be a world class institute – producing original work in policy and research.
- Leverage technology, networks and talent to extend our reach and influence.

PRIORITIES

- Initiate new programmes of work, that meet social, economic and environmental needs.
- Build digital capability and capacity to reach more people.
- Develop products and services that enable data sharing and use.
- Create more courses that build skills and knowledge with a wider customer base.
- Strive for equity in how we deliver what we do. Forge partnerships in new regions of the world, and among people who are under-represented.

COMMITMENTS

Vision
A world where data works for everyone

Mission
An open, trustworthy data ecosystem

IMPACT

The ODI's approach to this consultation

Since its inception, the Open Data Institute has been committed to our mission: to work with companies and governments to build an open, trustworthy data ecosystem. This is the ODI's response to the European Commission's open call for contributions on Competition in Generative AI.

Our response to this consultation builds on our extensive experience and thinking on open data and its impact on innovation and markets, as well as work from our Data-Centric AI programme.

At the ODI, we, along with other proponents of 'data-centric AI', discuss the essential role and dynamics of data for AI systems. As our co-founder Sir Nigel Shadbolt has told the UK parliament, "although we are talking a lot about AI, for the algorithms, their feedstock—their absolute requirement—is data".

As such, our response focuses in particular on issues associated with the role of data in generative AI systems. The 3 core aims of our data-centric AI programme are to:

- Make data AI-ready
- Make AI data accessible and usable
- Make AI systems use data responsibly

In this response, we discuss the importance of these core goals to competition for generative AI.

In this response, we answer questions 1, 2, 3, 4, 7, 8 and 10.

1) What are the main components (i.e., inputs) necessary to build, train, deploy and distribute generative AI systems? Please explain the importance of these components

As work from our data-centric AI programme has discussed, data, data infrastructure and data governance play an essential role throughout every stage of the AI lifecycle, from design to testing and ultimately delivery of (generative) AI-based services.

As mentioned in our introduction to this consultation, the ODI's approach is based on the notion that 'without data, there is no AI', and that data is fundamentally the 'feedstock' of models.

At the ODI, we typically talk about the 'data infrastructure' of AI to include data assets, tools, standards, data practices and data communities, all of which are essential to the role of data within systems. As we have elaborated upon in this December 2023 short report, forms of data used in generative AI tools are highly varied, with the training data used within a single model emerging from varying sources including web-crawling, enterprise data, or some combination of sources, and at times coming from millions of web domains. We highlighted quite different types of data included in generative AI models, which can fall into broad categories such as textual data, visual data and synthetic data. Contrary to assumptions that higher volumes of data or longer training time will result in improved outputs for generative AI systems, poor quality of data even in high volume is a serious risk for systems. As we discuss throughout this response, data is essential for the quality and relevance of model outputs, and data infrastructure is everything that ensures that data can continue to support effective data-enabled technologies and their required functions in society.

Data governance and stewardship, the processes overarching all of these elements of infrastructure, are essential for deployment and distribution of data in ways that ensure generative AI systems remain trustworthy. This includes transparency and accountability mechanisms to help understand and assess how data has been used in generative AI systems, something too often missing in the current moment, something ODI colleagues discussed in a Harvard Data Science Review article. As we discuss further in question 2, processes to address stakeholders' concerns associated with data used in systems (such as privacy and copyright concerns among various others) will be core to ensuring that generative AI systems can operate effectively.

The people involved in all parts of the lifecycle of generative AI systems are also crucial for ensuring effective systems that they can be responsibly operationalised. In particular, data literacy is key. For those in technical roles, this goes beyond technical knowledge and capability but also includes 'the ability to think critically about data in different contexts and examine the impact of different approaches when collecting, using and sharing data and information'. As part of this, it is essential to ensure that diverse needs are addressed in the creation and governance of generative AI systems. This means participation and user-centric design of these systems, particularly in relation to marginalised and minority groups.

2) What are the main barriers to entry and expansion for the provision, distribution or integration of generative AI systems and/or components, including AI models? Please indicate to which components they relate.

Barriers include:

- Concerns about data included in systems, which may lead to removal of public data and reduction in trust in generative systems, for example:
    - Copyright and intellectual property concerns fuelling lawsuits and protest across a range of industries, for example journalists and artists suing over the use of copyrighted material in training models.
    - Privacy concerns, such as whether personal data input into generative AI-based platforms is secure - a concern among others that has led many organisations to ban or limit the use of these platforms.

- Use of poor quality including biassed data, and lack of quality control over synthetic data - about which ODI co-founder and Chair, and Oxford University Professor of Computer Science Sir Nigel Shadbolt has publicly discussed significant concerns (for example in the Times and Evening Standard). Without quality interventions for data input, there is a risk of spiralling quality in model output - and ultimately 'model collapse', which "happens when generative AI becomes unstable, wholly unreliable or simply ceases to function" (December 2023 article by Sir Nigel Shadbolt). Serious risks of this type could degrade the quality of models, and trust long term.
- Costs of generating and/or accessing high quality datasets that are fit for purpose to generate effective models. Some organisations (such as big tech companies) have greater access to data to begin with – through their own systems and operations, or the ability to buy access to data – and the capability, skills and infrastructure required to make the most of it. Network effects intensify, as they link, aggregate and use the vast amount of data available to them – they continue to extract more and more value from their data.
- The costs for compute. In the UK, tackling this challenge is being supported by the central government. In their Autumn Statement, the UK government announced additional funding to support public compute, taking their planned investment to over £1.5billion. There was no additional funding in the budget, but there was a commitment to explaining how public compute facilities will be managed. It's hoped that, as a result, "both researchers and innovative companies are able to secure the computing power they need to develop world-class AI products".
- Data literacy among creators, deployers and users of models to address concerns about operationalising new systems, and ensuring they understand how to do so appropriately
- At the ODI, we have demonstrated the importance of trust in data ecosystems for supporting use of data and data-enabled systems, and therefore innovation and market benefits. As such, it is essential to address trustworthiness as a core need/potential barrier for generative AI systems
    - Including addressing any public concerns such as bias, discrimination, and other risks to safety and wellbeing.

3) What are the main drivers of competition (i.e., the elements that make a company a successful player) for the provision, distribution or integration of generative AI systems and/or components, including AI models?

Data access is key to supporting competition for organisations, as discussed in question 2. At the ODI we have long been strong proponents of high quality, *open* data for AI to increase innovation, and to ensure effective oversight of the data that is used in systems. This is on the basis that data can be 'non-rival' - meaning the same data can be used by many different systems and actors "at the same time without them needing to compete for access to it", so social and economic value can be derived from openness by many stakeholders. Barring the relevant compliance requirements, facilitating easy use of appropriate, high quality, open source datasets across the industry is essential.

Similar to the points made above, data skills and the ability to generate effective and trustworthy systems are also key, as public support will be essential for companies wanting to be successful in innovation and operationalising their systems.

Within our data-centric AI programme at the ODI, we have described the need to ensure data is 'AI-ready' so that successful systems can be implemented. This also includes ensuring competent stewardship approaches for data that address ethical and legal concerns including copyright issues and worker's rights, and standards and best practices that are clear and well-adhered to on a broader scale.

4) Which competition issues will likely emerge for the provision, distribution or integration of generative AI systems and/or components, including AI models? Please indicate to which components they relate.

Issues associated with the ownership and use of generative AI systems have already emerged. In question 2, we discussed issues associated with access to data, which we are aware are accompanied by significant challenges in terms of access to compute power. Development of monopolies can be associated with the ability to access data through existing practices of tech companies (e.g., including high volumes of information held as a product of their work) and acquisition of high quality data from smaller organisations (discussed in question 10). Evidence suggests that existing monopolies by big tech, largely in the US, in generative AI jobs are likely to continue "without intervention". There are also competition issues emerging in the geopolitical landscape, including in which nations have been able to be a part of the race to develop and shape these technologies.

The challenges are primarily associated with deconstructing pre-existing monopolies.

5) How will generative AI systems and/or components, including AI models, likely be monetised, and which components will likely capture most of this monetization?

-

6) Do open-source generative AI systems and/or components, including AI models compete effectively with proprietary AI generative systems and/or components? Please elaborate on your answer.

-

7) What is the role of data and what are its relevant characteristics for the provision of generative AI systems and/or components, including AI models?

We discuss this topic at length throughout this response. See in particular our answer to question 1.

8) What is the role of interoperability in the provision of generative AI systems and/or components, including AI models? Is the lack of interoperability between components a risk to effective competition?

In terms of data used within generative AI systems, there is huge value in interoperability between systems.

At the ODI, we are co-chairing a working group on the ML Commons machine learning data standard Croissant. As the ML Commons team points out, there is a need for such a standard for datasets "to make it easier to find and use ML datasets and especially to develop tools for creating, understanding, and improving ML datasets".

This reduces costs to identify and make use of high quality machine learning datasets, which will be increasingly important as broader efforts to ensure ML data improves develop. This will also support literacy and explainability about best practice in data practices for generative AI models - and facilitate accountability mechanisms to oversee and investigate how systems work.

At the ODI we have a long and broad history supporting open data initiatives across sectors to improve innovation and competition. For example our work on Open Banking as a member of the Open Banking Working Group that helped develop the standards and guidelines for open banking which has been hugely successful in enabling effective competition and innovation. Among other examples, our work on OpenActive has supported the creation of high quality open data in the sport and physical activity sector to help improve access to activities. These have reduced thresholds to make use of data across many different domains resulting in successful initiatives across the industries. We have undertaken extensive work on the importance of open data and open models to unlock social and economic benefits across the economy.

As we have spoken about earlier in this response, there are significant risks from lack of interoperability associated with accumulation of monopolies.

9) Do the vertically integrated companies, which provide several components along the value chain of generative AI systems (including user facing applications

and plug-ins), enjoy an advantage compared to other companies? Please elaborate on your answer.

-

10) What is the rationale of the investments and/or acquisitions of large companies in small providers of generative AI systems and/or components, including AI models? How will they affect competition?

As researchers from Epoch have demonstrated, there is a risk of “high quality” language data for AI ‘running out’ by 2026, and depletion/ultimate end of the supply of vision data, and even low-quality language data. This, they argue, will lead to stagnation of development of machine learning models. As such there may be increasing need/competition for high quality data.

This phenomenon may lead to the acquisition of smaller organisations by large ones pursuing the production of generative AI systems, because the stock of ‘publicly available’ data on the internet has already been used.

11) Do you expect the emergence of generative AI systems and/or components, including AI models to trigger the need to adapt EU legal antitrust concepts?

-

12) Do you expect the emergence of generative AI systems to trigger the need to adapt EU antitrust investigation tools and practices?

-