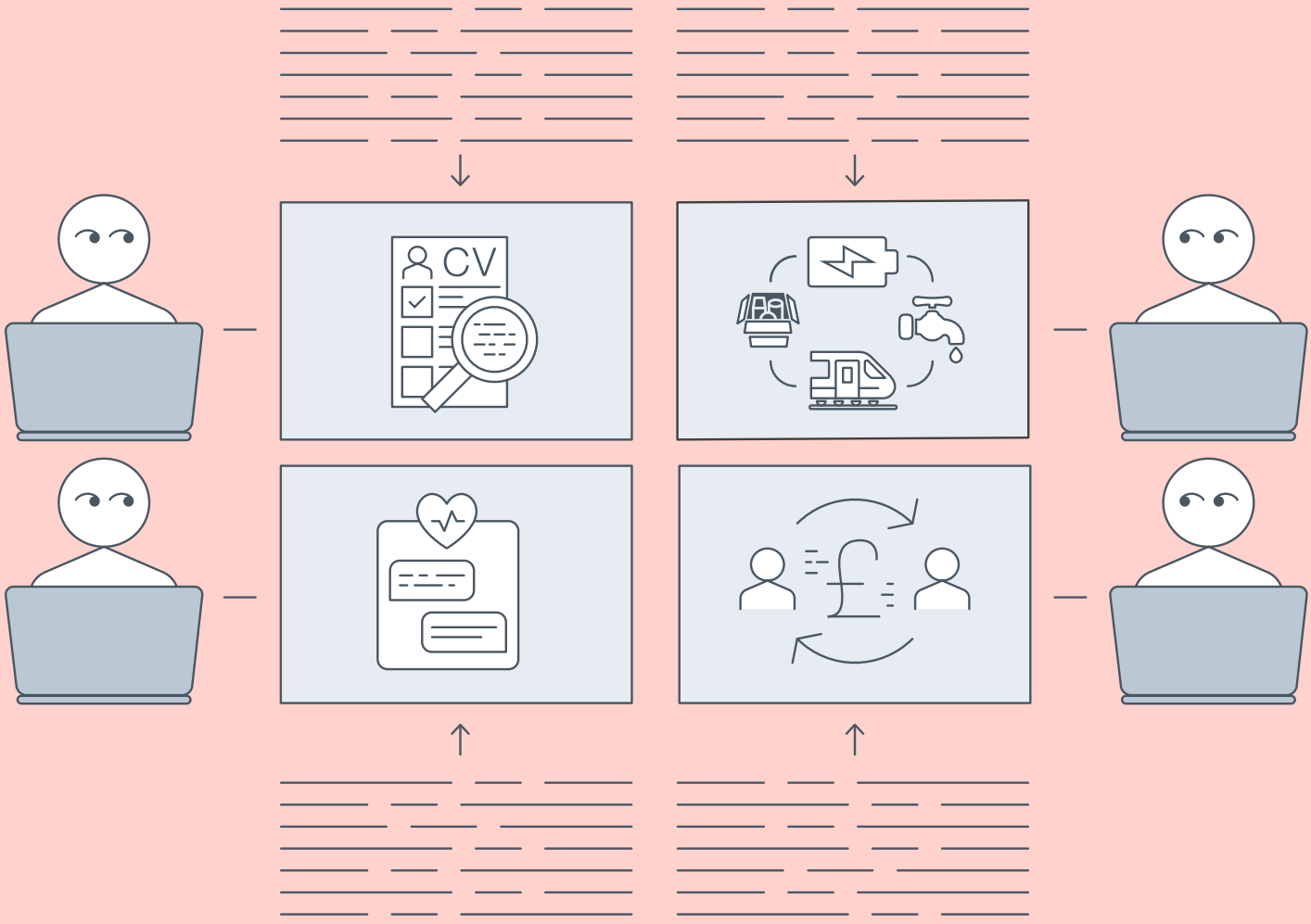
Ada Lovelace Institute

Law & policy

Keeping an eye on AI

Approaches to government monitoring of the AI landscape

July 2023



Contents

3 Executive summary

7 Introduction

12 How could Government monitor developments in AI?

19 What are the monitoring gaps in the existing ecosystem that Government is well placed to fill?

23 What mechanisms are there for sectoral and cross-cutting monitoring by individual regulators?

30 How can Government address the challenge of monitoring development and deployment of foundation models?

32 Conclusion

35 Further questions

37 Methodology

38 Partner information and acknowledgements

39 About the Ada Lovelace Institute

Just as governments closely monitor other sectors to inform policy, a similar approach can be adopted for AI.

Executive summary

The rapid development and deployment of artificial intelligence (AI) systems has the potential to be transformative for society. Whether its effects are beneficial or harmful, AI comes with an array of challenges that policymakers must navigate. To stay ahead of the curve and make well informed strategic decisions, it is essential that the UK Government possesses accurate and timely information about AI systems and their applications.

Just as governments closely monitor other sectors to inform policy, a similar approach can be adopted for AI. Consider how governments use inflation statistics to guide economic policymaking, or the creation of COVID-19 dashboards for public health decisions. These examples showcase the ability to aggregate complex data and distil it into actionable insights, which can then be used to shape effective policies.

In the context of AI, effective monitoring could provide crucial insights into the impacts, risks and opportunities of AI systems across different sectors and countries. By examining AI's influence on areas such as employment and recruitment, finance, healthcare and critical infrastructure, Government could make decisions that maximise benefits and mitigate potential injustices or harms. For instance, monitoring could reveal the need for new regulations to protect affected persons, or investments in safety research for foundation models.

Despite the clear need for comprehensive AI monitoring capabilities, the Government currently lacks the necessary infrastructure and tools to effectively track AI developments. By investing in robust AI monitoring systems, both Government and the public will be better equipped to make informed decisions, based on high-quality data, about whether to use or trust AI systems. This would create a deeper understanding of the societal impact of the use of AI systems and create a robust base for forward-thinking policies.

The key question of this paper is:

Given the speed of progress and complexity in AI development and deployment, how can Government source, rapidly synthesise and

summarise information about technological capabilities, trends, risks and opportunities in order to make informed, strategic policy decisions?

Key takeaways

1. There are specific properties of AI systems that the Government should consider measuring and monitoring such as their deployment and governance, and their downstream impacts on individuals and society. This information could include: inputs to AI systems; categorical information about the data and model underlying the AI systems; categorical information about processes or operations followed in development and deployment; direct outputs and outcomes of AI systems; and externalities generated by those outputs and outcomes.

2. The Government is well placed to address gaps in the existing monitoring ecosystem. This could be through standardised and mandated disclosure of information from companies, e.g. on compute; voluntary or statutory sharing of commercially sensitive information; and working with other governments on global comparative monitoring efforts. However, not all this information could or should be collected directly by central Government. Sectoral and cross-cutting regulators will have an essential role in providing contextual information during the gathering and interpretation of quantitative or aggregate data on AI capabilities and risks .

3. Regulators already have a number of existing mechanisms to ensure that information about the AI landscape is identified and shared with them. This information includes new developments, opportunities and risks with a sector or domain. Possible mechanisms are: standards for model cards and datasheets; regulatory sandboxes or multi-agency advisory services, regulatory inspection and audit; whistleblower protections and rewards; incident reporting; and ombudsmen.

4. Foundation models present unique challenges for cross-sectoral regulation. If and when AI applications in different sectors become more reliant on foundation models, it will not be efficient for individual regulators to each individually assess and monitor these systems and create multiple, overlapping and potentially conflicting demands

on the companies deploying them. One option to address this issue is the creation of a centralised AI regulatory function with institutional relationships with the relevant sectoral or cross-cutting regulators and direct monitoring relationships with developers of foundation models.

5. The UK Government should not delay in building out its internal monitoring capabilities, and should immediately initiate small, focused pilot projects responding to policy challenges faced by the Government. This could include establishing a national-level public repository of the harms, failures and unintended negative consequences of AI systems; building on the Review of the Future of Compute; putting in place Government monitoring, aggregating (and potentially publication) of data on broad compute use and demand trends; and requesting to be informed when frontier AI labs begin large-scale training runs of new models.

Research questions

This paper aims to understand how Government can better measure and monitor developments in the AI landscape, addressing the following research questions:

1. What value can monitoring AI developments provide?

2. What aspects of AI research, development and deployment could be measured?

3. How could Government monitor the AI landscape?

4. What are the monitoring gaps in the existing ecosystem that the Government is well placed to fill?

5. What mechanisms are there for sectoral and cross-cutting monitoring by individual regulators?

6. How can the Government address the challenges of monitoring across development and deployment of AI foundation models?

This paper provides an analysis of approaches to monitoring developments in the AI landscape and outlines elements of AI systems

that Government could monitor. It then analyses existing private and intergovernmental initiatives for systematically monitoring AI development and deployment, and identifies several gaps that the Government is uniquely be able to fill. It then examines approaches to monitoring and responsible disclosure in individual sectors, including sandboxes, incident reporting and horizon scanning. Finally, this paper concludes with a discussion of how these approaches could be complicated by the introduction of general-purpose AI systems and future questions that need to be explored.

Key terms

Measurement is the collection of information that reduces (expected) uncertainty about a given topic. This requires deciding what information needs to be collected and how to best gather or collect it.1

Monitoring is the process of operationalising measurement over time and entails systemically and continually gathering measurements in a common format.2 Monitoring allows for tracking changes over time and for that information to be aggregated and integrated in policymaking.

Ex ante mechanisms are forward-looking regulatory tools that take effect before an AI system is deployed and impacts users and affected people. Examples of ex-ante mechanisms include regulatory sandboxes which allow companies and regulators to test AI products in a real-world environment before they enter the market.

Ex post mechanisms are backwards-looking regulatory tools that take effect after an AI system is deployed. These include regulatory inspection and auditing methods, in which an AI system is evaluated and tested by a regulator for compliance with a particular law or regulation.

1 Jess Whittlestone and Jack Clark, 'Why and How Governments Should Monitor AI Development' (arXiv, 31 August 2021) 5 http://arxiv.org/abs/2108.12427 accessed 7 February 2023.

2 ibid 6.

Introduction

The rapid development and deployment of artificial intelligence (AI) systems present both opportunities and challenges for governments and society. Given the speed and complexity of AI progress, it is crucial for governments to have high quality and timely information about AI systems and their uses in order to make informed, strategic policy decisions.

This paper aims to understand how the Government can better measure and monitor developments in the AI landscape, addressing the following research questions:

1. What value can monitoring AI developments provide?
2. What aspects of the AI research, development and deployment could be measured?
3. How could the Government monitor the AI landscape?
4. What are the monitoring gaps in the existing ecosystem that the Government is well placed to fill?
5. What mechanisms are there for sectoral and cross-cutting monitoring by individual regulators?
6. How can the Government address the challenges of monitoring across development and deployment of AI foundation models?

Drawing from academic and grey literature (including preprints) on government AI monitoring, this paper analyses potential approaches to measuring and monitoring AI developments. It examines case studies from the private sector and academia to explore how the Government might approach AI monitoring, and identifies gaps that the Government is uniquely placed to fill. The paper also investigates mechanisms for sector-specific monitoring and discusses how insights from individual sectors can be aggregated to inform Government policymaking.

By developing robust AI monitoring capabilities, the Government may be able to better anticipate and understand AI's impact on society and mitigate potential inequalities or harms

What value can monitoring AI developments provide?

Governments already monitor developments and changes in different sectors for the purposes of policymaking. Inflation statistics are a prime example of how governments aggregate high-quality information to inform strategic policy-making. By distilling information from thousands of prices into a single estimate, governments enhance public understanding of the economic situation and can more effectively shape monetary and fiscal policy. Governments also undertake similar practices for monitoring epidemiological and global health developments, such as the creation of COVID-19 dashboards.[3] While creating these statistics is no easy task and requires significant resources, governments and the public recognise their value in creating an evidence base for policymaking, making it easier to anticipate and govern emerging challenges.[4]

Effective AI monitoring could similarly distil a complex set of information from across different sectors and countries to provide insights into the impacts, risks and opportunities of AI systems. However, the UK Government lacks such monitoring capabilities for AI developments. By developing robust AI monitoring capabilities, the Government may be able to better anticipate and understand AI's impact on society, identify opportunities for investment, and mitigate potential inequalities or harms.[5] This could enable both the public and the Government to make informed decisions based on high-quality data.

In addition to supporting compliance and impact assessment, monitoring AI developments is essential for the Government to better understand its AI ecosystem relative to other countries. This knowledge can help prioritise funding for research and innovation, ensuring that the nation remains competitive in the global AI landscape. A thorough understanding of research, development and deployment across the AI value chain, both domestically and internationally, is vital for maintaining a strong position in an increasingly technology-driven world.

3 Whittlestone and Clark (n 1).

4 As Baron Alexandre Lamfalussy, President of the European Monetary Institute, put it in 1996: "Nothing is more important for monetary policy than good statistics. Statistical information is necessary to decide what policy actions to take, to explain them publicly, and to assess their effects after the event. Unless policy can be justified and explained, it will not be understood and the institution carrying it out will lack credibility." Europäische Zentralbank (ed), *Statistics and Their Use for Monetary and Economic Policy-Making: Second ECB Conference on Statistics*, 22 and 23 April 2004 (European Central Bank 2004) 40.

5 Whittlestone and Clark (n 1).

It is crucial for the Government to have its own monitoring system to ensure accurate and unbiased insights into how AI is being researched, developed and deployed

Monitoring AI developments also addresses the potential issue of information asymmetries between the Government and the private sector. If the Government does not actively monitor AI developments, it may lead to a higher incidence of AI accidents and misuse, and deployments with negative externalities, followed by hasty and uninformed legislation.

Furthermore, relying solely on external providers for AI monitoring creates dependencies that may leave the Government vulnerable. Commercial interests and the priorities of developers may drive the provision of available information, which could result in a skewed understanding of the domestic AI ecosystem and lead to suboptimal policy decisions. For example, OpenAI did not disclose detailed information on the architecture (including model size); hardware; training compute; dataset construction; training method or similar in its technical report for its flagship large language model (LLM), GPT-4. The reason given for withholding information was 'the competitive landscape and the safety implications of large-scale models'.[6]

While supporting external monitoring efforts is important for experimentation and niche areas, it is crucial for the Government to have its own monitoring system to ensure accurate, unbiased and comprehensive insights into how AI is being researched, developed and deployed, domestically and internationally. This will empower the Government to make well-informed decisions based on a solid foundation of knowledge.

Lastly, monitoring AI developments can help mitigate 'race' dynamics in AI development.[7] By investing in intelligence on the state of AI capabilities in other countries, governments can better understand the global landscape and avoid unnecessary competition to deploy powerful AI systems. Knowledge of other countries' progress in AI – whether they are close to deploying potentially harmful systems or not – informs strategic decisions and ensures a more responsible approach to AI

6 OpenAI, 'GPT-4 Technical Report' (OpenAI 2023) 2 https://cdn.openai.com/papers/gpt-4.pdf accessed 16 March 2023. There are good reasons why a private company may choose not disclose this information, for example to prevent competitors replicating the model, or to slow the spread of capabilities. Nevertheless, this leaves the decision to share information about cutting-edge AI capabilities in the hands of private companies, making it harder for the government to assess the appropriate policy response and for regulators and other third-parties to assess the capabilities and potential societal effects of these models.

7 Holden Karnofsky, 'How Major Governments Can Help with the Most Important Century' (*Cold Takes*, 24 February 2023) https://www.cold-takes.com/how-governments-can-help-with-the-most-important-century/ accessed 16 March 2023.

development and deployment.

Once this monitoring and identification process is underway, transparency and information sharing should be encouraged between governments. For example, public indexes could be set up that measure and display key metrics for safe AI deployment (e.g. fairness, or lack thereof, in AI systems, with the index ranking the 'most fair' systems).

This would benefit both consumers, who would have better information about products, and companies, who would be incentivised to compete with each other to develop and deploy systems that top such rankings. It would also disincentivise companies to rush to deploy potentially harmful systems (given the reputational issues from being measured and ranked on that basis). The Government should work with industry, civil society and affected persons (those who ultimately use or are affected by AI) to decide the metrics that are measured and indexed.

What aspects of AI research, development and deployment could be measured?

There are a number of properties of AI systems, their deployment and governance, and their downstream impacts on individuals and society that the Government could consider measuring and monitoring. These include:

Inputs to AI systems such as data, software, hardware, compute, knowledge, time or energy.[8] These could provide insights into the resource requirements of AI systems and the availability of those resources, and could inform quantitative assessments of cross-country AI readiness.

Categorical information about the data and model of the AI systems, such as the identity and location of the developer and/or deployer, the number of parameters, modalities, model architecture, cost, etc.

Categorical information about processes or operations followed in development and deployment, such as benchmarks used in tests (and

8 For more detail on potential inputs to AI systems, see the table in Fernando Martínez-Plumed and others, 'Between Progress and Potential Impact of AI: The Neglected Dimensions' (arXiv, 2 July 2022) 4 http://arxiv.org/abs/1806.00610 accessed 8 February 2023.

performance on those benchmarks), impact assessments, oversight mechanisms such as safety checklists or ethics committees. This information could allow regulators and civil society to better scrutinise how a system was tested, what criteria or process was used, and the results of those tests, and also whether a particular issue was the result of poor testing or governance.

Direct outputs and outcomes of AI systems. Outputs and outcomes could include the number of users served, number of tokens generated, revenue generated, or number of documented accidents and misuse of the system.

Externalities generated by the development and deployment of AI systems, for example, environmental footprint, infringement of user and affected persons' privacy, and skill atrophy among users. Measurement of these externalities is an important step towards internalising them, for example by requiring companies to pay the full carbon cost of developing their models, and thus incentivising them towards developing more energy efficient systems.[9]

Overall, there is a wide range of different indicators and evidence the Government might wish to measure in order to have a holistic picture of technological capabilities, trends, risks and opportunities to support informed, strategic policy decisions.

9 ibid 2.)

How could the Government monitor developments in AI?

There are current attempts by global governmental bodies to monitor AI systems that we can learn from. The European Commission's Joint Research Centre's AI Watch has found in its own attempts at monitoring that private performance benchmarks, competitions and challenges are behind much of the recent progress in AI.[10] However, it also found that these metrics alone do not provide an appropriate way for policymakers and other stakeholders to assess what AI systems can do today and in the future, in a way that can inform appropriate policy.

The private sector frequently evaluates the performance of AI systems using specific, quantitative benchmarks that seek to provide comparative measures on aspects like accuracy, precision, recall and processing speed. These are sometimes complemented with more qualitative measures of performance over relatively short time horizons, such as user feedback, evaluations by users directly in the product (e.g. thumbs-up or thumbs-down on a recommendation or piece of generated text), user reports within the product, and even aggregating comments on the product posted to social media.

Although some companies publicise their AI performance on benchmark leaderboards, often explicitly competing to surpass the state of the art, this information is not consistently disclosed to the public.[11] These are potentially important indicators of the capabilities of AI systems, but they are currently not functional for that purpose. Governments could engage with and shape benchmarking development efforts and encourage these to be made public routinely (with clear acknowledgements of their limitations to mitigate unwarranted hype).

10 Fernando Martınez-Plumed, Jose Hernandez-Orallo and Emilia Gomez, 'Tracking the Impact and Evolution of AI: The AIcollaboratory' 1.

11 ibid; Percy Liang and others, 'Holistic Evaluation of Language Models' (arXiv, 16 November 2022) http://arxiv.org/abs/2211.09110 accessed 29 March 2023; Aarohi Srivastava and others, 'Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models' (arXiv, 10 June 2022) http://arxiv.org/abs/2206.04615 accessed 29 March 2023.

In terms of measurement and standardisation, most benchmarks address performance metrics relating to accuracy and recall that can be quantitatively measured. However, there is a lack of consistent benchmark metrics and tests for issues that address wider societal implications of these systems, including standardised tests for bias or fairness issues.[12] There are also no standardised processes for reporting qualitative or reflexive information about the potential risks or impacts of these technologies, who will be impacted by them, the likelihood of potential impacts, and what steps a company has taken to address or mitigate those risks.

These benchmarks do not give the whole picture; they are a measure of performance, but not how that performance was achieved. For example, systems might achieve high scores that do not generalise outside the benchmark, if testing data is accidentally or intentionally included in the training dataset by the developers. These benchmarks also do not provide an assessment of how to reproduce the evaluated performance, and therefore how to verify the claims made by developers. These benchmarks therefore do not incentivise a rigorous approach to measuring AI performance that will be essential for the creation of a thriving market of AI products. Without that rigour and reproducibility, it is hard to create a system of assurance around products that are safe, reliable and do what they say on the tin.

These benchmarks also have limited scope. They often do not measure the societal benefits or impacts of these technologies. They are measuring performance on a discrete set of tasks, but often lack an evaluation of how these performance metrics translate into impacts on different parts of society. Basing policy decisions purely on these benchmark metrics could lead to a misallocation of resources, overinvesting into tools that are unreproducible, unexplainable and do not serve society's interests. A more holistic view of AI progress could be given through other measures such as compute efficiency, data efficiency, novelty, replicability, explainability and ability for the system to generalise performance beyond its training dataset.[13]

12 Inioluwa Deborah Raji and others, 'AI and the Everything in the Whole Wide World Benchmark' (arXiv, 26 November 2021) http://arxiv.org/abs/2111.15366 accessed 26 March 2023; Michelle Bao and others, 'It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks' (arXiv, 28 April 2022) http://arxiv.org/abs/2106.05498 accessed 26 March 2023.

13 Martínez-Plumed and others (n 8) 1.

A key indicator that the Government could measure (and that some private initiatives are already attempting to catalogue)[14] is compute access and usage. OpenAI Researchers found that there is a direct relationship between the amount of computational power used in the development of large language models (and similar systems) and their potential to have downstream impacts, for better or worse.[15] Monitoring where large amounts of compute are being used would therefore allow governments to develop early awareness of which actors are likely to be developing and deploying highly capable (and risky) systems.

Monitoring does already exist: there are a number of organisations and initiatives that aim to systematically catalogue and monitor the emerging AI landscape with a broader view than just performance benchmarks. These range from overarching summaries of the field to specialised indexes and catalogues for particular dimensions of the landscape, like compute or documented cases of accidents.[16]

However, these approaches have limitations. Some are discontinued or published by individuals or small non-profit companies that are not guaranteed to continue the work in future, or take a very narrow view of inputs or properties of the systems. Some of these approaches rely primarily on crowdsourced data, such as the AI Incidents Database or the Metaculus forecasts, and will be skewed by the biases of those contributors. They are also reliant on unpaid volunteers for the continued upkeep of the monitoring effort. Previous surveys of initiatives to monitor AI progress have found similar limitations.[17]

In the table below we summarise a (non-comprehensive) survey of exiting initiatives, highlighting their organisational structure, cadence, what they cover and how they measure it, and respective strengths and weaknesses as monitoring approaches.

14 Jaime Sevilla and others, 'Parameter, Compute and Data Trends in Machine Learning' https://docs.google.com/spreadsheets/d/1AAIebjNsnJj_uKALHbXNfn3_YsT6sHXtCU0q7OlPuc4/edit accessed 16 March 2023.

15 Jared Kaplan and others, 'Scaling Laws for Neural Language Models' (arXiv, 22 January 2020) http://arxiv.org/abs/2001.08361 accessed 23 March 2023."plainCitation":"Jared Kaplan and others, 'Scaling Laws for Neural Language Models' (arXiv, 22 January 2020

16 Ian Hogarth and Nathan Benaich, 'State of AI Report 2022' (2022) https://www.stateof.ai/ accessed 16 March 2023; Daniel Zhang and others, 'The AI Index 2022 Annual Report', (AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University 2022) https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf accessed 16 March 2023; Responsible AI Collective, 'Welcome to the Artificial Intelligence Incident Database' (2023) https://incidentdatabase.ai/ accessed 16 March 2023.

17 Martınez-Plumed, Hernandez-Orallo and Gomez (n 10) 2.

AI measurement and monitoring in practice

State of AI Report[18]

The *State of AI Report*, created by Nathan Benaich and Ian Hogarth, is an annual publication that offers a detailed analysis of the previous year's trends and developments in artificial intelligence across research, industry, talent and policy.

- **Cadence:** Annual
- **Type of organisation:** Venture capitalist
- **What do they measure? How do they measure them?**
  - Overview of AI landscape.
  - Over the course of the year, the authors note what they view as the most significant developments in AI research, industry, safety and politics, summarising particular examples and then outlining the trends they have seen over the previous year.
  - They occasionally draw on funding, investment and market share data to illustrate their points.
  - Ultimately, the authors compile and analyse what they deem to be the most interesting things they have seen over the previous year.
- **Strengths**
  - Wide if not comprehensive coverage. These reports provide a thorough analysis of various aspects of AI, including research, industry applications, talent distribution and policy considerations, offering readers a holistic understanding of the AI landscape.
  - Accessibly written. The reports are structured and presented in an easily digestible manner, making complex information about AI accessible to a wide range of readers, including policymakers.
- **Weaknesses**
  - Given the fast-paced nature of AI development, some information in the reports may become outdated relatively quickly, waiting until the following year to be updated.
  - As with any analysis, the *State of AI Report* may be subject to the perspectives and biases of their authors, in this case venture capitalists, potentially skewing the presentation of their findings to reflect industry concerns.

AI Index[19]

The *AI Index* is an annual report produced by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) that tracks and analyses AI's progress across various dimensions, including research, technology development, investments, policy and societal impact.

- **Cadence:** Annual
- **Type of organisation:** Academic
- **What do they measure? How do they measure them?**
  - Overview of AI landscape.
  - The Index team compiled data from existing research, including CSET, Emsi Burning Glass and LinkedIn (employment and skills data).
  - They created original datasets on benchmark performance and use of metrics compiled from publicly available papers.
  - For AI-related bills and mentions in legislatures, the AI Index performed searches of the keyword 'artificial intelligence', in the respective languages, on the websites of 25 countries' congresses or parliaments.
  - They provide a full methodology of all data collection in the report appendix
- **Strengths**
  - Openly provides the raw data used for the report and an interactive tool for cross-country comparisons.
- **Weaknesses**
  - Given the fast-paced nature of AI development, some information in the reports may become outdated relatively quickly, waiting until the following year to be updated.

18 Hogarth and Benaich (n 16).

19 Zhang and others (n 16).

State of AI Report Compute Index[20]

Nathan Benaich and Ian Hogarth also compile a *State of AI Report* Compute Index, that tracks the size of public, private and national high-performance computing (HPC) clusters, as well as the utilisation of various AI chips in AI research papers.

- **Cadence:** ~Monthly (but not consistently)
- **Type of organisation:** Venture capitalist
- **What do they measure? How do they measure them?**
  - Number of NVIDIA A100 graphics processor units (GPUs) available to tech companies and public-sector compute clusters.
  - Chip type usage in open-source AI papers that cite the use of specific AI chips, broken down into NVIDIA chips, start-up chips, and other chips (e.g. Google's in-house tensor processor units (TPUs).
  - Systematic monitoring of data, compute, cost and parameters of large models.
  - The Epoch staff compiled existing work cataloguing large models. They then added information from publicly available evidence they deemed significant or representative of a given year, back to 1950.
  - Where data, e.g. on compute or cost, is not published, the staff team instead provide best-guess estimates based on proxies that are available and inferences from the data that is available.
- **Strengths**
  - Offers a clear comparison in terms of access to compute by public and private actors.
  - Systematic comparison: Epoch's dataset, with the standardised data across hundreds of models, allows for comparison across different AI systems and the identification of trends in compute use, modality, cost, etc.
- **Weaknesses**
  - Reliant on publicly declared access to and usage of specific chips.
  - Underlying data not easy to extract and reuse.
  - Infrequently updated and maintained.
  - Narrow in scope, even in domain of compute.
  - Epoch's approach is reliant on the accuracy of the data that it uses, and the accuracy of their extrapolations when it comes to compute and cost.
  - Epoch is reliant on developers of large models openly publishing information about their systems, which means it will miss unreleased models and those without public documentation (especially as frontier labs become less open) and the database will be backward looking.

Parameter, Compute and Data Trends in Machine Learning Database[21]

Epoch research group aims to forecast the development of transformative AI by gathering information about the timing of new developments, studying which factors influence AI progress and examining current and past trends in ML. They have compiled a public dataset of over 500 notable and / or large AI models developed since 1950 to aid in tracking trends.[22]

- **Cadence:** Live, with reports published
- **Type of organisation:** Non-profit
- **What do they measure? How do they measure them?**
  - Systematic monitoring of data, compute, cost and parameters of large models.
  - The Epoch staff compiled existing work cataloguing large models. They then added information from publicly available evidence they deemed significant or representative of a given year, back to 1950.
  - Where data, e.g. on compute or cost, is not published, the staff team instead provide best-guess estimates based on proxies that are available and inferences from the data that is available.

20 Ian Hogarth and Nathan Benaich, 'State of AI Report Compute Index' (2023) https://www.stateof.ai/compute accessed 16 March 2023.

21 The Epoch Team, 'Epoch Impact Report 2022' (Epoch, 1 February 2023) https://epochai.org/blog/epoch-impact-report-2022 accessed 22 March 2023.

22 Sevilla and others (n 14).

- **Strengths**
  - Offers a clear comparison in terms of access to compute by public and private actors.
  - Systematic comparison: Epoch's dataset, with the standardised data across hundreds of models, allows for comparison across different AI systems and the identification of trends in compute use, modality, cost, etc.
- **Weaknesses**
  - Epoch's approach is reliant on the accuracy of the data that it uses, and the accuracy of their extrapolations when it comes to compute and cost.
  - Epoch is reliant on developers of large models openly publishing information about their systems, which means it will miss unreleased models and those without public documentation (especially as frontier labs become less open) and the database will be backward looking.

Responsible AI Collaborative's AI Incident Database[23]

The AI Incident Database collection of publicly reported instances where AI systems have caused unintended negative consequences or failures is curated by the Responsible AI Collaborative. It aims to raise awareness about the risks and challenges associated with AI deployment by documenting real-world cases across various domains.

- **Cadence:** Whenever new reports are received
- **Type of organisation:** Non-profit
- **What do they measure? How do they measure them?**
  - News reports involve AI accidents and misuse.
  - Incident reports are submitted by individual users
  - Reports are then accepted or rejected by their editorial team on the basis of an evolving set of rules defined in their Editor's Guide.
- **Strengths**
  - Provides a structured, accessible repository of AI incidents. The database serves as a valuable resource for researchers, developers, policymakers, and other stakeholders to learn from past mistakes and identify patterns of system failure.
- **Weaknesses**
  - Reliant on news and user reports, so will likely skew towards high-profile and public failures of AI systems, rather than low-level but pervasive flaws.

Metaculus AI and Machine Learning Forecasts[24]

Metaculus online forecasting platform and aggregation engine maintains records for thousands of forecasters and generates aggregate forecasts on forecasted questions. Questions of AI progress and governance are a popular topic on Metaculus and they have run forecasting tournaments specifically focused on AI,[25] alongside hiring a team specifically focused on AI forecasting.[26]

- **Cadence:** Continuously updated
- **Type of organisation:** Public benefit corporation
- **What do they measure? How do they measure them?**
  - Crowdsourced and aggregate forecasts of questions on AI development, deployment and governance e.g.:
  - 'Will Google or DeepMind release a public interface for a large language model before April 1, 2023?'
  - 'When will a Chinese organisation train a large model on GPUs that deliver more than 1500 TFLOPs@FP16?'
  - 'Will Chase Bank send a notice to customers about updates to its security protocols, referencing the threat of social engineering attacks that can be attributed to LLMs, before 2024?'

---

23 Responsible AI Collective (n 16).

24 Metaculus, 'About' https://ai.metaculus.com/about/ accessed 22 March 2023.

25 Metaculus, 'Forecasting AI Progress' https://www.metaculus.com/project/ai-progress/ accessed 22 March 2023.

26 Christian Williams, 'Metaculus Is Building a Team Dedicated to AI Forecasting' https://www.lesswrong.com/posts/JqtPkFqmJtoHTahis/metaculus-is-building-a-team-dedicated-to-ai-forecasting accessed 22 March 2023.

- **Strengths**
  - Draws on the wisdom of the crowds and rapid aggregation of relevant information by forecasters
  - Changes in probability, especially sharp ones, are a signal there is new relevant information to consider.
- **Weaknesses**
  - Some forecasts have few entries and so likely high error.
  - Expertise of forecasters unclear and may have incentives other than accuracy, e.g. pushing their own beliefs about the world.
  - Analysis has previously found that Metaculus forecasters were biased towards thinking AI things will happen sooner than they actually did, and that they were particularly overconfident about numeric questions, e.g. what score will the best system get on a specific benchmark on a specific date.[27]

AI Watch[28]

AI Watch is a project of the European Commission Joint Research Centre. It aims to be knowledge service to monitor the development, uptake and impact of AI in Europe. It monitors:

  - AI Enablers – the role of data and infrastructure in powering AI systems
  - AI Landscape – overview and analysis of the AI landscape
  - AI Standards
  - Evolution of AI technology – analysing the evolution of AI technology, methodologies, breakthroughs and benchmarks
  - Trustworthy AI – analysing the impact of AI on people and society.
- **Cadence:**
  - Index published annually.
  - Case studies and reports published frequently, but not on a specific schedule.
  - Dashboards are not based on live data.
- **Type of organisation:** Intergovernmental organisation
- **What do they measure? How do they measure them?**
  - Interactive dashboards of AI strategies, the AI landscape and AI investment across EU countries.
  - Case studies on particular applications of AI, e.g. facial recognition and autonomous vehicles.
  - An AI Watch Index, [29] that analyses (i) global view on the AI landscape, (ii) industry, (iii) research and development (R&D), (iv) technology, and (v) societal aspects.
- **Strengths**
  - AI Watch covers various aspects of AI development, including enablers, landscape, standards, technology evolution and societal impact. This comprehensive approach provides a well-rounded understanding of AI in Europe.
  - Being a project of the European Commission Joint Research Centre ensures that AI Watch is developing outputs optimised for policymakers, ensuring that its findings are useful by stakeholders.
- **Weaknesses**
  - AI Watch focuses primarily on AI development in Europe, which may not provide a complete understanding of global AI trends and advancements.
  - The dashboards are not based on live data, which means that they may not always reflect the most current state of AI development.

---

27 Charles Dillon, 'An Examination of Metaculus' Resolved AI Predictions and Their Implications for AI Timelines' (Rethink Priorities, 20 July 2021) https://rethinkpriorities.org/publications/an-examination-of-metaculus-resolved-ai-predictions accessed 22 March 2023.

28 Martınez-Plumed, Hernandez-Orallo and Gomez (n 10) 1.

29 Riccardo Righi and others, 'AI Watch Index 2021' (JRC Publications Repository, 30 March 2022) https://publications.jrc.ec.europa.eu/repository/handle/JRC128744 accessed 21 March 2023.

What are the monitoring gaps in the existing ecosystem that the Government is well placed to fill?

Compared to private companies, the Government possesses unique powers to access direct, continuous and high-value information for monitoring the AI landscape. For example, Government can obtain commercially sensitive information, systematic reports and audit results through their status as a trusted intermediary and through statutory powers if necessary. This access could enable the Government to collect and analyse data that is unavailable to private organisations or academic researchers.

For example, many attributes of AI systems are not systematically reported. These include quantitative data such as compute used for training and deployment, or information about the size, modalities and provenance of training data.. Private monitoring initiatives, like Epoch discussed above, often rely on the voluntary disclosure of information by companies and inferences based on technical evaluations of AI models.[30]

However, Government could mandate standardised reporting and aggregation of these attributes by developers and researchers, including requiring the use of datasheets and model cards that document this information. Government could also require that cloud compute providers support this process, for example by requiring prospective customers to create and provide appropriate documentation as a condition of use for training models and sharing information on compute usage.[31]

30 Dario Amodei and Danny Hernandez, 'AI and Compute' (OpenAI, 16 May 2018) https://openai.com/blog/ai-and-compute/ accessed 26 March 2020; Jaime Sevilla and others, 'Estimating Training Compute of Deep Learning Models' (Epoch, 20 January 2022) https://epochai.org/blog/estimating-training-compute accessed 17 March 2023.

31 Noam Kolt, 'Algorithmic Black Swans' (25 February 2023) 48 https://papers.ssrn.com/abstract=4370566 accessed 23 March 2023."plainCitation":"Noam Kolt, 'Algorithmic Black Swans' (25 February 2023

These practices could create a more reliable assessment of the global competitive landscape with regard to compute and AI research, and help identify emerging trends and potential risks of emerging AI technologies. They could also reverse the tendency of developers to keep cutting-edge capabilities and understanding of those capabilities private from Government and regulators, due to concerns about information hazards,[32] and a desire to keep a competitive edge, through proprietary intellectual property and trade secrets.[33] Ultimately, a democratically representative government has the legitimacy to decide what information is hazardous and what information it needs to best regulate increasingly capable AI systems.

Sometimes there is a good case for this information not being widely disclosed, for example sharing datasets that contain highly sensitive personal data, or sharing model architecture and weights for powerful models that increase the capabilities of malicious actors. For example, enabling 'spear-phishing' fraud,[34] or personalised manipulation and harassment at a previously cost-prohibitive scale.[35]

In these instances, the Government could still gain insight into the development of these systems through voluntary disclosure via regulatory sandboxes or multi-agency advisory services. The Government could help companies better understand the societal implications of their cutting-edge systems and provide regulatory guidance in exchange for forewarning about the capabilities and deployment of these systems. If that information is not forthcoming voluntarily, the Government could also require that companies and developers have a statutory obligation to disclose information of systems that meet a certain risk threshold, e.g. compute required or particular capabilities in high-risk domains.

32 An information hazard is a risk that arises from the dissemination or the potential dissemination of (true) information that may cause harm or enable some agent to cause harm. Nick Bostrom, 'Information Hazards' (2011) 10 Review of Contemporary Philosophy 44, 2. In this context, it refers to the potential diffusion of information about how to develop advanced AI systems that would allow malicious actors, e.g. an authoritarian government, to deploy capabilities they would not otherwise have access to, e.g. the ability to generate hyper-personalised propaganda.

33 Ada Lovelace Institute, 'Regulate to Innovate' (2021) 43 https://www.adalovelaceinstitute.org/report/regulate-innovate/ accessed 1 February 2023; Toby Shevlane, 'Sharing Powerful AI Models | GovAI Blog' (20 January 2022) https://www.governance.ai/post/sharing-powerful-ai-models accessed 16 March 2023; OpenAI (n 6) 2.

34 A more targeted variation of phishing attacks, where usually the victim is addressed by name and is often otherwise personalised to the victim to make the fraudulent activity more convincing. See: Tian Lin and others, 'Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content' (2019) 26 ACM Transactions on Computer-Human Interaction 1.

35 Josh A Goldstein and others, 'Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations' (arXiv, 10 January 2023) 41 http://arxiv.org/abs/2301.04246 accessed 24 March 2023.

The Government should take more global view of the AI landscape when discussing monitoring approaches

It is important to note that not all this information could or should be collected directly by the Government. Sectoral and cross-cutting regulators will have an essential role in providing contextual information which is crucial in gathering and interpreting quantitative or aggregate data on AI capabilities and risks. By leveraging their unique expertise and understanding of specific sectors, regulators can contribute valuable insights and perspectives to the overall picture of the AI landscape, ensuring a more nuanced and well-informed approach to policy development.

For example, in the context of automated diagnostic tools, the Medicines & Healthcare products Regulatory Agency (MHRA) and Care Quality Commission (CQC) have existing relationships with specialised diagnostic device providers. These provide an understanding of the evaluation and monitoring of existing diagnostic tools and their integration into the wider healthcare system.[36] The MHRA and CQC are therefore better placed to gather information about the development and deployment of automated diagnostics, and place them in the wider context of healthcare for policymakers.

Many frontier AI capabilities are being developed outside the UK, and the initial instances of many benefits and harms may occur elsewhere. It is therefore necessary to take a more global view of the AI landscape when discussing monitoring approaches. The Government may find it beneficial to consider international developments in AI to ensure a comprehensive understanding of emerging trends, risks and opportunities in the rapidly evolving AI domain. Efforts like the OECD AI Observatory provide an example of an attempt at this kind of monitoring. At the moment, much of the OECD monitoring focuses on AI strategies, skills and investment, but it is also exploring a framework for understanding, measuring and benchmarking domestic AI computing supply by country and region.[37]

36 Care Quality Commission, 'Using Machine Learning in Diagnostic Services: A Report with Recommendations from CQC's Regulatory Sandbox' (March 2020) https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf accessed 26 March 2023.

37 'The OECD Artificial Intelligence Policy Observatory' (OECD, 2023) https://oecd.ai/en/ accessed 26 March 2023; 'The OECD.AI Expert Group on AI Compute and Climate' (OECD, 2020) https://oecd.ai/en/p/compute accessed 26 March 2023.

However, sectoral monitoring mechanisms should adopt a more UK-focused lens, as the unique information-gathering powers of the Government and regulators are often limited to the bounds of UK jurisdiction. By addressing these gaps in the AI monitoring ecosystem, Government can effectively support responsible AI development and informed policymaking both nationally and internationally.

What mechanisms are there for sectoral and cross-cutting monitoring by individual regulators?

In individual sectors, regulators also need an understanding of the applications of AI in their respective areas of oversight. There are a number of existing mechanisms that could be used to ensure that information about the AI ecosystem (including new developments, opportunities and risk with a sector or domain) is identified, while minimising the compromise of commercial integrity or generation of new risks from knowledge diffusion.

Regulators are well placed to gather direct, on-the-ground information about their sectors through both ex ante and ex post monitoring mechanisms, informing both their own decision-making and the Government's policies. Ex ante mechanisms are forward-looking regulatory tools that aim to operate before AI is deployed and exposed to users and affected people, e.g. regulatory sandboxes for pre-market AI products. Ex post mechanisms are backwards-looking regulatory tools that operate after AI is deployed, e.g. regulatory inspection and audit of AI systems already in production or incident reporting for misuse and accidents.

Below, we outline a number of these mechanisms, highlighting existing or proposed examples of each approach, and their strengths and weaknesses.

Mechanisms for regulators to measure and monitor AI

Standards for model cards[38] and datasheets

Model cards and datasheets are tools used to document an AI system's characteristics, training data, limitations, intended use, and potential biases and flaws. Implementing sector-specific standards for these tools could enhance transparency, accountability and compliance with regulatory requirements. It would also make comparison between systems easier to undertake systematically.

- ***Ex ante*** or ***ex post***?
  - Ex ante
  - Existing examples
  - Model cards are in use by Google,[39] Hugging Face[40] and others, and other organisations are developing comparable approaches, such IBM's AI Factsheets.[41]
- Strengths
  - Improved transparency, better understanding of AI systems' limitations.
  - Easier comparison across different AI models.
  - Facilitates better-informed decision-making for AI system users and regulators.
- Weaknesses and limitations
  - Increased documentation workload for developers and potential disclosure of proprietary information.
  - May not cover all relevant aspects of an AI system and may be challenging to enforce across an entire sector, especially on companies who primarily operate outside the jurisdiction.

Regulatory sandboxes[42]

Regulatory sandboxes allow companies to test innovative products, services and business models in a controlled environment, sometimes with temporary regulatory exemptions. This enables regulators to assess new technologies, identify potential risks and develop appropriate regulatory responses.

- ***Ex ante*** or ***ex post***?
  - Ex ante
- Existing examples
  - The UK Financial Conduct Authority (FCA)'s Regulatory Sandbox and Bank of England's FinTech Accelerator.
  - The Information Commissioner's Office (ICO) has an AI sandbox for data protection and privacy.
  - The Norwegian Data Protection Agency introduced an AI regulatory sandbox following a British model.[43]

38 Margaret Mitchell and others, 'Model Cards for Model Reporting', Proceedings of the Conference on Fairness, Accountability, and Transparency (2019) http://arxiv.org/abs/1810.03993 accessed 1 February 2023.

39 Google Cloud, 'Google Cloud Model Cards' (2023) https://modelcards.withgoogle.com/about accessed 17 March 2023.

40 Ezi Ozoani, Marissa Gerchick and Margaret Mitchell, 'Model Cards' (20 December 2022) https://huggingface.co/blog/model-cards accessed 17 March 2023.

41 M Arnold and others, 'FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity' (2019) 63 IBM Journal of Research and Development 6:1. IBM Research, 'AI FactSheets 360' (2023) https://aifs360.mybluemix.net/examples accessed 17 March 2023.

42 Lea Maria Siering and Till Christopher Otto, 'Regulatory Sandboxes' (Lexology, 5 February 2020) https://www.lexology.com/library/detail.aspx?g=419b7b84-bde0-4c29-bb63-41df2aa3d0b1 accessed 21 March 2023.

43 Birgitte Kofod Olsen, 'Sandbox For Responsible Artificial Intelligence · Dataetisk Tænkehandletank' (Dataetisk Tænkehandletank, 14 December 2020) https://dataethics.eu/sandbox-for-responsible-artificial-intelligence/ accessed 21 March 2023.

  - Articles 53-55 of the EU AI Act propose AI regulatory sandbox set up by a national competent authority to develop, train, validate and test, where appropriate in real-world conditions, innovative AI systems, according to a specific plan for a limited time under regulatory supervision.[44]

- **Strengths**
  - Provides insight into pre-market AI systems and supports more informed policy decisions.[45]
  - Regulated companies have access to early regulatory insight and are able to make more informed product development decisions, reducing regulatory burden on start-ups in a controlled environment.[46]
  - A dedicated outreach and engagement strategy is needed to succeed and can be costly but is important for knowledge gathering. For example, The Consumer Financial Protection Bureau had a successful Office Hours programme, leading to many fintech companies voluntarily sharing product information.
- **Weaknesses and limitations**
  - Resource-intensive for regulators, even at the design stage. The development of a regulatory sandbox in one developing economy took 18 months and even in high-income states the development of a regulatory sandbox typically requires a minimum of at least six months.[47]
  - 'Almost two thirds of those regulators interviewed [by the UNSGSA] noted that they had significantly underestimated the resources required to develop and operate their sandboxes.'[48]
  - These costs limit the ability for regulators to scale regulatory sandboxes without additional resources from government to cover additional dedicated staff.
  - Sandbox participants may not be representative of the entire industry, and findings may not be generalisable. For example, the EU AI Act's focus on start-ups and small-scale providers will miss insights for larger companies.
  - While sandboxes can provide information on the effects of a single product in a controlled setting, it is unlikely to provide information on the market or systemic effects of a new AI product.[49]
  - Risks of regulatory arbitrage, where regulators lower safeguards and requirements to get insights and promote innovation.[50]

Multi-agency advisory services

Multi-agency advisory services offer support to start-ups navigating the regulatory landscape, enabling the Government to identify upcoming AI technologies with significant regulatory implications.

- ***Ex ante* or *ex post*?**
  - *Ex ante*
- **Existing examples**
  - AI and Digital Regulations Service for health and social care, created by the National Institute for Health and Care Excellence (NICE), the Medicines and Healthcare products Regulatory Agency (MHRA), the Health Research Authority (HRA), and the Care Quality Commission (CQC). [51]

44 Permanent Representatives Committee (Part 1) to Council of the European Union, 'Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts - General Approach' (25 November 2022) https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf accessed 21 March 2023.

45 Dan Quan, 'A Few Thoughts on Regulatory Sandboxes' (Stanford PACS, 2019) https://pacscenter.stanford.edu/a-few-thoughts-on-regulatory-sandboxes/ accessed 21 March 2023.

46 Siering and Otto (n 47).

47 UNSGSA FinTech Working Group and CCAF, 'Early Lessons on Regulatory Innovations to Enable Inclusive FinTech: Innovation Offces, Regulatory Sandboxes, and RegTech' (Offce of the UNSGSA and CCAF 2019) 31 https://www.unsgsa.org/sites/default/files/resources-files/2020-09/UNSGSA_Report_2019_Final-compressed.pdf accessed 21 March 2023.

48 ibid.

49 Sofia Ranchordas, 'Experimental Regulations for AI: Sandboxes for Morals and Mores' (SSRN, 12 July 2021) 16 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3839744&download=yes accessed 19 June 2023.

50 Hilary J Allen, 'Sandbox Boundaries' (2020) 22 Vanderbilt Journal of Entertainment and Technology Law 299, 315.

51 NHS AI and Digital Regulations Service for health and social care, 'Regulatory Guidance of Digital Technology in Health and Social Care - AI Regulation Service - NHS' (2023) https://www.digitalregulations.innovation.nhs.uk/ accessed 21 March 2023.

- **Strengths**
  - Streamlines regulatory compliance and identifies potential risks and opportunities associated with new AI technologies.
  - Enables proactive identification of regulatory challenges and promotes collaboration between start-ups and regulators.
- **Weaknesses and limitations**
  - Resource-intensive for regulators and may create dependencies between start-ups and Government agencies.
  - The advisory services' effectiveness depends on the level of cooperation and information-sharing between participating agencies.

Regulatory inspection and audit

Regulators having statutory powers to investigate and test AI systems for monitoring, suspected noncompliance or verifying claims.

Conducting a meaningful regulatory inspection of an algorithmic system would require regulators to have powers to accumulate specific types of evidence, including information on:

  - categorical information about processes or operations followed in development and deployment, including both policies i.e. company policies and documentation that identify the goals of the AI system, what it seeks to achieve, and where its potential weaknesses lie; and processes i.e. assessment of a company's process for creating the system, including what methods they chose and what evaluation metrics they have applied.
  - the outputs and outcomes of AI systems on a range of different users of the system.
- ***Ex ante* or *ex post*?**
  - *Ex post*
- **Existing examples**
  - In legislation on algorithms on social media platforms in the EU's Digital Services Act or UK's Online Safety Bill.[52]
- **Strengths**
  - Audits usually come into place after a system is in use, so can serve as accountability mechanisms to verify whether a system behaves as developers intend or claim, whether risk mitigations have been effective and investigate whether unanticipated impacts have occurred.
  - A survey of auditors almost unanimously reported that their largest barriers are lack of buy-in from auditees to conduct audits in the first place, and limited enforcement capabilities. Auditors often want to disclose their results and methods, but are restricted by nondisclosure agreements.[53] Regulators undertaking statutory audits can overcome all these barriers with appropriate legislative powers and uncover information about the real-world impact of AI systems that private and non-governmental auditors and monitors are unable to access.
- **Weaknesses and limitations**
  - While approaches are emerging, there is no single source for best practices, which make it more challenging for regulators to adopt methods without upfront investment developing their own tools and techniques.[54]
  - The number of AI developers and deployers means that regulators undertaking audit of more than a fraction of systems is likely to be unfeasible. However, audit action can still be an effective information gathering tool when focused on systems deemed high-risk, utilising frontier capabilities that are not yet generally well understood, or are otherwise opaque to other forms of evidence gathering.

---

52 Ada Lovelace Institute, 'Technical Methods for Regulatory Inspection of Algorithms in Social Media Platforms' (Ada Lovelace Institute 2021) https://www.adalovelaceinstitute.org/wp-content/uploads/2021/12/ADA_Technical-methods-regulatory-inspection_report.pdf accessed 1 February 2023.

53 Sasha Costanza-Chock, Inioluwa Deborah Raji and Joy Buolamwini, 'Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem', 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM 2022) 8–9 https://dl.acm.org/doi/10.1145/3531146.3533213 accessed 17 March 2023.

54 ibid 6.

Whistleblower protections

Whistleblower protections encourage individuals to report wrongdoing within their organisation by providing legal and financial safeguards

- ***Ex ante* or *ex post*?**
  - *Ex post*
- **Existing examples**
  - In the UK, the Public Interest Disclosure Act (1998) protects individuals who 'blow the whistle' in the public interest. [55] This covers a wide range of sectors, from financial malpractice or risks to health to environmental damage. The disclosure must usually be made to an appropriate external body, e.g. the Health and Safety Executive or the Financial Conduct Authority.
  - Similar protections exist in, for example the European Union[56] and Australia.[57]
- **Strengths**
  - Empowers individuals to report concerns without fear of retaliation.
  - Unveils unethical or harmful practices that may otherwise remain hidden.
  - Promotes a culture of transparency and accountability.
- **Weaknesses and limitations**
  - Possibility of false or misleading reports.
  - May not be sufficient to protect whistleblowers in all cases.
  - Relies on individuals coming forward, which may not happen consistently.

Whistleblower rewards

Whistleblower reward programmes incentivise whistleblowers to come forward by giving them some portion of any fine successfully prosecuted. A reward scheme could also offer flat or scaling rewards not linked directly to prosecutions and fines.

- ***Ex ante* or *ex post*?**
  - *Ex post*
- **Existing examples**
  - The use of these schemes in the UK so far is somewhat limited, including up to £100,000 for reporting on illegal cartel activity to the Competition and Markets Authority and an HMRC reward scheme for individuals who reported tax fraud, including fraud related to the COVID-19 relief schemes.[58]
  - In the USA, under the Motor Vehicle Safety Whistleblower Act, whistleblowers can receive a reward of up to 30% of any monetary sanctions up to $1,000,000 dollars for exposing safety-related problems
- **Strengths**
  - Creates a proactive channel to reduce information asymmetries between government and industry.
  - Incentivises even staff not motivated primarily by safety or misuse concerns to flag malpractice and proactively share information with regulators by rewarding them with a portion of any fine levied.
  - May incentivise AI labs to be more cautious, and provide a signal to staff that their concerns will be listened to by policymakers.
- **Weaknesses and limitations**
  - Investigations often take many years and hard-to-justify payouts before a successful investigation, which potentially limits any upside for whistleblowers.[59]

55 Public Interest Disclosure Act 1998.
56 Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons who report breaches of Union law 2019.
57 Public Interest Disclosure Act 2013.
58 Protect, 'Whistleblowing – a Rewarding Act?' (21 April 2022) https://protect-advice.org.uk/whistleblowing-a-rewarding-act/ accessed 22 March 2023.
59 Protect, (2022).

- This could create perverse incentives for whistleblowers to hold off on raising concerns until there is a bigger fine. It also creates an incentive to bypass the employer's internal whistleblowing arrangements.
- As of 2014, the FCA and Prudential Regulation Authority claimed that there is as yet no empirical evidence of incentives leading to an increase in the number or quality of disclosures received by the regulators.[60]

Incident reporting to regulators

Incident reporting mechanisms require organisations to notify regulators of specific events, such as data protection breaches, that may have significant consequences. This allows regulators to assess the situation and take appropriate action.

In an anonymous survey of 152 individuals who engage in algorithmic audits or whose

work is directly relevant to algorithmic audits, establishment of systematic harm incident reporting was the third-highest priority (of 13) identified for regulatory intervention.[61]

- ***Ex ante* or *ex post*?**
  - *Ex post*
- **Existing examples**
  - The General Data Protection Regulation (GDPR) requires companies to report data breaches to relevant authorities within 72 hours.[62]
  - Many countries operate national systems for the notification of occupational injuries, supplemented by surveys of households and employers. For example, this is done by the Health and Safety Executive in the UK[63] or the Occupational Safety and Health Administration in the USA.[64] These statistics are then aggregated by the International Labour Organisation to provide cross-country comparisons.[65]
- **Strengths**
  - Enables prompt regulatory intervention. Encourages a culture of accountability and transparency.
  - Helps regulators identify patterns and trends in AI-related incidents.
- **Weaknesses and limitations**
  - Under-reporting may be an issue due to fear of penalties or reputational damage.
  - Changes in reporting standards over time can limit the ability for comparisons of trends across time.

Ombudsman

An ombudsman is an independent authority that investigates complaints by consumers and citizens. This could take the form of empowering existing ombudsmen or creating a new ombudsman specifically for AI systems and their deployment, ensuring impartiality and fairness.

- ***Ex ante* or *ex post*?**
  - *Ex post*
- **Existing examples**
  - Housing Ombudsman
  - Pension Ombudsman

60 Financial Conduct Authority and Prudential Regulation Authority to Treasury Select Committee, 'Financial Incentives for Whistleblowers' (July 2014) https://www.fca.org.uk/publication/financial-incentives-for-whistleblowers.pdf.

61 Costanza-Chock, Raji and Buolamwini (n 59) 6.

62 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) 2016 art 33.

63 The Reporting of Injuries, Diseases and Dangerous Occurrences Regulations 2013.

64 Occupational Safety and Health Administration, 'Recordkeeping - Overview' (2023) https://www.osha.gov/recordkeeping/ accessed 22 March 2023.

65 ILOSTAT, 'Occupational Safety and Health Statistics (OSH Database)' (ILOSTAT, 2023) https://ilostat.ilo.org/resources/concepts-and-definitions/description-occupational-safety-and-health-statistics/ accessed 22 March 2023.

  - The Swedish Equality Ombudsman investigated how a state-owned bank used an algorithm to calculate individual credit risk for those aged over 60 years old in 2018, leading to the bank changing its rules.[66]
- **Strengths**
  - Offers an independent and impartial avenue for addressing AI-related grievances.
  - Provides a systematic route to collecting data on certain kinds of AI accidents and misuse.
- **Weaknesses and limitations**
  - A new ombudsman would require resources for establishing and maintaining the office. Other major sectoral ombudsman schemes in the UK have budgets from the low £10s of millions to around £250 million for the Financial Ombudsman Service.[67]
  - May lack enforcement power and existing ombudsmen have noted difficulties interrogating automated decision-making systems.[68]
  - Dependent on the scope and mandate given by the Government.

66 David Coulter, 'Regulating for an Equal AI: A New Role for Equality Bodies' (Equinet 2020) https://equineteurope.org/wp-content/uploads/2020/06/ai_guide_digital.pdf.

67 Financial Ombudsman Service, (2022), Our 2023/24 plans and budget, Consultation paper, p. 15, https://www.financial-ombudsman.org.uk/files/324119/Financial-Ombudsman-Service-Plans-and-Budget-Consultation-2023-24.pdf (Accessed: 29 March 2023) Legal Ombudsman, (2022), Business Plan 2022/23, p. 34, https://www.legalombudsman.org.uk/media/l22jss5j/20220324-2022-23-olc-business-plan-and-budget-final.pdf (Accessed: 29 March 2023); The Housing Ombudsman, (2023), Annual Report and Accounts 2021-22, p. 64, https://www.housing-ombudsman.org.uk/wp-content/uploads/2023/02/E02841575-Housing-Ombudsman-ARA-21-22_elay.pdf (Accessed: 29 March 2023); The Pensions Ombudsman, (2022), Corporate Plan 2022-2025, p. 30, https://www.pensions-ombudsman.org.uk/sites/default/files/publication/files/TPO%20Corporate%20Plan%202022-25_0.pdf (Accessed: 29 March 2023).

68 Ombudsperson, Province of British Columbia and others, 'Getting Ahead of the Curve: Meeting the Challenges to Privacy and Fairness Arising from the Use of Artificial Intelligence in the Public Sector' (2021) Joint Special Report 2 25.

The rapid development of foundation models means that there is even more urgency for a monitoring approach that acknowledges the shift happening in the development and deployment of AI systems

How can the Government address the challenge of monitoring development and deployment of foundation models?

Foundation models are AI systems are designed to produce a wide and general variety of outputs. They are capable of a range of possible tasks and applications, such as text synthesis, image manipulation, or audio generation.[69] Sometimes called general-purpose AI (or GPAI), they can be used in standalone systems or as the 'building block' of hundreds of single-purpose AI systems to accomplish a range of distinct tasks, such as sector-specific analytic services, e-commerce chatbots or designing a custom curriculum – often without substantial modification and fine-tuning.

Current foundation models are characterised by their scale, using huge datasets featuring billions of words or hundreds of millions of images scraped from the internet and millions of pounds worth of compute per training run, and their reliance on transfer learning (applying knowledge from one task to another) – although future foundation models may not necessarily have these properties.[70] In contrast, narrow AI applications are those trained for a specific task and context, that are difficult to reuse for new contexts.

We are increasingly seeing the deployment of foundation models, and many of these systems are made accessible via an application programming interface (API), where developers allow users to use and potentially fine-tune – but not fundamentally modify – the underlying

69 Sabrina Küspert, Nicolas Moës and Connor Dunlop, 'The Value Chain of General-Purpose AI' (Ada Lovelace Institute Blog, 10 February 2023) https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/ accessed 27 March 2023.

70 Risto Uuk, 'General Purpose AI and the AI Act' (Future of Life Institute 2022) https://artificialintelligenceact.eu/wp-content/uploads/2022/05/General-Purpose-AI-and-the-AI-Act.pdf accessed 26 March 2023.

system. Many downstream providers in turn build context-specific AI applications on top of this foundation-model API. For example, with the launch of OpenAI's GPT-4, we've seen many companies start to build products underpinned by the GPT-4 API, including Microsoft's BingChat,[71] Duolingo Max,[72] Khan Academy's Khanmigo[73] or Be My Eyes' Virtual Volunteer.[74]

The rapid development of foundation models means that there is even more urgency for a monitoring approach that acknowledges the shift happening in the development and deployment of AI systems. If and when narrow AI applications become more reliant on foundation models, it will not be efficient for individual regulators to individually assess and monitor these systems and create multiple, overlapping and potentially conflicting demands on the companies deploying them.

This issue could be addressed by the creation of a centralised AI regulatory function that establishes direct monitoring relationships with developers of foundation models. It could act as a coordinating body within Government, channelling concerns and information from individual regulators about downstream applications that rely on upstream foundation models. It could then establish where there were recurring issues across different sectoral contexts as a result of the design or implementation of upstream foundation models. It could also manage subsequent requests for information, collating responses from foundation-model providers and sharing that information across Government and regulators.

In a UK context, this could take the form of establishing a centralised body with institutional relationships with relevant regulators and empowering it to establish direct relationships with companies.

71 Yusuf Mehdi, 'Reinventing Search with a New AI-Powered Microsoft Bing and Edge, Your Copilot for the Web' (The Official Microsoft Blog, 7 February 2023) https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ accessed 26 March 2023.

72 Duolingo Team, 'Introducing Duolingo Max, a Learning Experience Powered by GPT-4' (Duolingo Blog, 14 March 2023) https://blog.duolingo.com/duolingo-max/ accessed 26 March 2023.

73 Sal Khan, 'HArnessing GPT-4 so That All Students Benefit. A Nonprofit Approach for Equal Access!' (Khan Academy Blog, 14 March 2023) https://blog.khanacademy.org/harnessing-ai-so-that-all-students-benefit-a-nonprofit-approach-for-equal-access/ accessed 26 March 2023.14 March 2023

74 Be My Eyes, 'Introducing Our Virtual Volunteer Tool for People Who Are Blind or Have Low Vision, Powered by OpenAI's GPT-4' (2023) https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer accessed 26 March 2023.

Conclusion

In its March 2023 white paper on AI regulation, *A pro-innovation approach to AI regulation*, the UK Government proposed creating a set of central Government functions to support the work of regulators.[75] Some of these functions could be used to carry out the monitoring activities outlined in this paper:

- **Horizon scanning:** a function intended to directly monitor emerging trends and opportunities in the AI development landscape, to support Government decision-making.
- **Cross-sectoral risk assessment:** a function intended to monitor known risks, and (in conjunction with the horizon-scanning function) identify and prioritise new risks.
- **Support for innovators:** a function that can gather valuable information on the state of AI development and what products are on the horizon via voluntary information sharing and evaluations in testbeds and sandboxes.

However, the first iteration of all the central functions are not due to be established until more than 12 months after the publication of the AI regulation white paper. Given the fast pace of AI development, it's crucial that the Government doesn't delay in building its internal monitoring capacity. Government needs systematic monitoring of the development AI systems and the risks they can pose now – to inform and iterate its approaches to future regulation, and ground discussions with evidence of actual harms and benefits.

Starting small with pilot projects for monitoring can build the necessary expertise and infrastructure for larger monitoring projects. To ensure the success of these pilots, they should have a clear scope and realistic ambitions, respond to specific policy challenges faced by Government policy stakeholders, and be provided with sufficient funding and staffing to achieve their ambitions.

75 Department for Science, Innovation & Technology and Office for Artificial Intelligence, 'A Pro-Innovation Approach to AI Regulation' (2023) https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper accessed 16 June 2023.

Potential immediate pilots could include:

- Establishing a national-level public repository of the harms, failures and unintended negative consequences of AI systems that are being deployed in the real world, and potential risks of in-development applications. This initiative could build on the work of the Responsible AI Collaborative's AI Incident Database, but in a more systematic fashion that draws on incidents reported by existing regulators that may not be public.[76] This kind of pilot could set the groundwork for monitoring future risks by more complex AI systems, which many developers of AI systems have called for. For example, 96% of respondents in a survey of 51 experts (from frontier AI labs, and those working in civil society and academia on frontier AI governance) agreed that frontier AI labs should report safety incidents and near misses to appropriate state actors. [77]

- Start regularly monitoring, aggregating (and potentially publishing) data on broad compute use and expectations of future compute requirements for model training and inference. This would build on the work of the Independent Review of the Future of Compute.[78] This would require leveraging information from financial reporting, import duties, export controls, and information volunteered by AI companies and researchers in government-run foresight exercises. This information could be complemented by the data already aggregated by organisations like Epoch, and could form the basis for reporting requirements for users and providers of large-scale compute for training AI models. As discussed above, compute is a critical aspect for AI progress (at least for current powerful frontier models – it is possible this may change over time) and much more easily monitored than other inputs such as data or talent. Beginning to collect and act on information about compute usage will make it easier to identify major AI developers and deployers systematically, and is currently a promising proxy for high-risk capabilities, allowing the Government to more effectively direct regulatory attention and risk-assessment

76 Responsible AI Collective (n 16).

77 Jonas Schuett and others, 'Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion' 5 https://www.governance.ai/research-paper/towards-best-practices-in-agi-safety-and-governance.

78 Zoubin Ghahramani and others, 'Independent Review of The Future of Compute: Final Report and Recommendations' (2023) https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts accessed 16 June 2023.

to those capabilities.[79] However, in future, increasing compute efficiency, a shift towards more specialised models or changes in model architecture could lead to compute being a less useful proxy for capabilities.

- The Government should request to be informed when organisations plan to run large-scale training runs of new AI models. This proposal builds on Google DeepMind, OpenAI and Anthropic's announced commitment to give early or priority access to models for research and safety purposes.[80] This access could provide Government with an early warning of advancements in frontier capabilities, allowing policymakers and regulators to prepare for the impact of these developments. It remains to be seen how much time Government will need to prepare appropriately and what steps they can take to respond. Nikhil Mulani and Jess Whittlestone have outlined in further detail what a voluntary pilot for monitoring model capability evaluations and compute requirements of foundational models developed by frontier labs could look like in practice.[81]

By commencing a pilot and starting to collect and act on the information, Government can better identify high-risk capabilities of AI systems and direct regulatory attention where it's needed most.

79 Jess Whittlestone and others, 'Future of Compute Review - Submission of Evidence' (Centre for Long-Term Resilience 2022) 9–10 https://www.longtermresilience.org/post/future-of-compute-review-submission-of-evidence accessed 16 June 2023.

80 'PM London Tech Week Speech: 12 June 2023' (GOV.UK, 12 June 2023) https://www.gov.uk/government/speeches/pm-london-tech-week-speech-12-june-2023 accessed 16 June 2023.

81 Nikhil Mulani and Jess Whittlestone, 'Proposing a Foundation Model Information-Sharing Regime for the UK' (GovAI Blog, 16 June 2023) https://www.governance.ai/post/proposing-a-foundation-model-information-sharing-regime-for-the-uk accessed 16 June 2023.

Further questions

How can insights from individual sectors be aggregated and used to inform Government policymaking? While we have made some progress in this paper towards answering this question, there are still open questions about how much data gathering and sharing can be standardised across sectors, and how best to inform policymakers whose work is directly affected by the use of AI systems but who are not part of a central coordinating function.

What statutory powers will Government need to undertake effective AI monitoring? As the Government considers putting its AI regulatory framework on a statutory footing in the future, it will be important to consider whether existing information-gathering powers and voluntary disclosure are sufficient, or whether regulators will need to be further empowered to undertake auditing or require standardised information disclosure from companies.

What other methods (e.g. horizon scanning) can Government undertake for AI monitoring? How effective are these methods? In order to be proactive, monitoring and subsequent policymaking could be complemented by methods that extrapolate out trends and provide more qualitative, in-depth pictures of the future.

What are the gaps and blindspots that sectoral regulators may have when it comes to monitoring AI developments? How can these be identified? As we discuss in the final chapter of this paper, cross-cutting AI foundation-models could increasingly underpin everyday AI applications and fall through the cracks between different sectoral regulators.

What are reliable and rigorous ways for monitoring the impacts of AI on labour and skills developments? There have been many reports examining the 'potential' impact of AI on labour and skills, but to establish whether those predictions are being borne out and the actual pace and scale of any labour market disruptions, Government and bodies like the Office for National Statistics (ONS) are best place to monitor that information.[82]

82 PwC, 'The Potential Impact of Artificial Intelligence on UK Employment and the Demand for Skills' (2021) BEIS Research Report 2021/042 https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1023590/impact-of-ai-on-jobs.pdf accessed 29 March 2023; Tyna Eloundou and others, 'GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models' (arXiv, 23 March 2023) http://arxiv.org/abs/2303.10130 accessed 29 March 2023.

Methodology

The methodology for this report involved a review of academic and grey literature pertaining to how governments can monitor AI developments. This began with the following research questions, which were translated into keyword searches on Google Scholar and a research assistance tool, Elicit, that plugs into Semantic Scholar:

- How [Could | Should | Can | Do] governments monitor AI development?
- What are [systematic] approaches to monitoring AI development?
- What are the limitations of current approaches to monitoring AI development?
- What are historical examples of [monitoring | disclosure | knowledge-sharing] that could be applied to AI?

Based on the results of initial searches, we used a snowballing method of searching that looked at papers referenced by papers that appeared in initial searches. We then searched for other papers by authors of papers in the initial searches, which were deemed relevant to the literature review.

Partner information and acknowledgements

This report was authored by Elliot Jones, with substantive contributions from Jenny Brennan, Connor Dunlop and Andrew Strait.

This work was undertaken with support via UKRI by the Department for Digital, Culture, Media & Sport (DCMS) Science and Analysis R&D Programme. It was developed and produced according to UKRI's initial hypotheses and output requests. Any primary research, subsequent findings or recommendations do not represent DCMS views or policy and are produced according to academic ethics, quality assurance and independence.

About the Ada Lovelace Institute

The Ada Lovelace Institute was established by the Nuffield Foundation in early 2018, in collaboration with the Alan Turing Institute, the Royal Society, the British Academy, the Royal Statistical Society, the Wellcome Trust, Luminate, techUK and the Nuffield Council on Bioethics.

The mission of the Ada Lovelace Institute is to ensure that data and AI work for people and society. We believe that a world where data and AI work for people and society is a world in which the opportunities, benefits and privileges generated by data and AI are justly and equitably distributed and experienced.

We recognise the power asymmetries that exist in ethical and legal debates around the development of data-driven technologies, and will represent people in those conversations. We focus not on the types of technologies we want to build, but on the types of societies we want to build.

Through research, policy and practice, we aim to ensure that the transformative power of data and AI is used and harnessed in ways that maximise social wellbeing and put technology at the service of humanity.

We are funded by the Nuffield Foundation, an independent charitable trust with a mission to advance social well-being. The Foundation funds research that informs social policy, primarily in education, welfare and justice. It also provides opportunities for young people to develop skills and confidence in STEM and research. In addition to the Ada Lovelace Institute, the Foundation is also the founder and co-funder of the Nuffield Council on Bioethics and the Nuffield Family Justice Observatory.

Find out more:

Website: Adalovelaceinstitute.org
Twitter: @AdaLovelaceInst
Email: hello@adalovelaceinstitute.org

Permission to share: This document is published under a creative commons licence: CC-BY-4.0

Preferred citation: Ada Lovelace Institute. *Keeping an eye on AI: Approaches to government monitoring of the AI landscape* (2023). Available at: https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai

ISBN: 978-1-7392615-3-5