Ada Lovelace Institute

Emerging Technology & Industry Practice

Under the radar?

Examining the evaluation of foundation models

July 2024

Contents

Executive summary

Global policy proposals for ensuring the safety of advanced artificial intelligence (AI) systems have centred on foundation model evaluations as an important method to identify and mitigate the risks these systems pose. The core goals of foundation model evaluations are to understand the foundation model and / or its impacts, including the model's capabilities, risks, performance, behaviour and social impact.

Policymakers are seeking to use evaluations to provide clarity on appropriate and responsible uses of foundation models. They are incorporating evaluations into emerging regulatory proposals in the EU, UK and USA, and creating both voluntary and legally mandated requirements for developers to evaluate AI systems for different kinds of risks.

The EU's newly passed AI Act requires developers of foundation models and general-purpose AI models to evaluate these systems for 'systemic risks'. The Act has established an AI Office, which also has a mandate to evaluate general purpose AI models.

In the USA and UK, governments have secured voluntary commitments from major AI companies to allow external evaluations of their foundation models by newly established national AI safety institutes. France, Canada, Japan and Singapore have their own AI safety institutes with similar mandates to develop and run evaluations of foundation models.

Both governments and technology companies have described evaluations as a necessary component of effective foundation model governance. Many foundation model developers have hired dedicated evaluation teams to construct evaluations and test their models, and there is also a growing third-party evaluation industry in which contracted third parties can construct test models on behalf of a developer.

However, our research indicates that evaluations alone are not sufficient for determining the safety of foundation models, the systems built from them and their applications for people and society in real-world conditions. There is no agreed terminology or set of methods for evaluating foundation models, and evaluations need to be used alongside other tools including codes of practice, incident reporting and post-market monitoring. In practice, AI model evaluations are currently voluntary and subject to company discretion, leading to inconsistencies in quality and limited access for evaluators without pre-existing company relationships. Current policy proposals allow companies to selectively choose what evaluations to conduct, and fail to ensure evaluation results lead to meaningful action that prevents unsafe products from entering the market.

What are foundation models?

Foundation models, sometimes called a 'general-purpose AI' or 'GPAI' system, are capable of a range of general tasks (such as text synthesis, image manipulation and audio generation).1 Notable families of foundation models are Google's Gemini 1, Anthropic's Claude 3 and OpenAI's GPT-4. The latter underpins the conversational chat agent ChatGPT and many other applications via OpenAI's application programming interface (API).

Foundation models are designed to work across many complex tasks and domains, and can exhibit complex, unpredictable and contradictory behaviour when prompted by human users. Unlike other industries like aerospace and medicine, which utilise strong theoretical underpinnings to prove the generality and validity of safety tests, the theoretical understanding of foundation models is currently lacking.

As these technologies are capable of a wide range of general tasks, they differ from narrow AI systems (those that focus on a specific or limited task, for example, predictive text or image recognition) in important respects:

- It can be harder to identify and foresee the ways foundation models can benefit people and society.
- It is harder to predict how foundation models may be used, in what contexts they will be deployed, how they can affect end users and other people impacted by the system, and therefore when they can cause harm.
- Foundation models' failures could have systemic, cascading effects, if hundreds of applications depend on a single foundation model.

It is important to note that what is accessible to users, and to those building on top of foundation models, is often a 'filtered' foundation model that

[1] To learn more about foundation models, see: 'What Is a Foundation Model?' (*Ada Lovelace Institute*) https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

includes additional components (see Figure 1. These components enable or disable certain kinds of interactions with the underlying model. The additional components may enable new kinds of behaviour or change the distribution of outputs for a given set of inputs, compared with an unfiltered model.

There are several general challenges for foundation model evaluations, with most falling into three categories. Firstly, challenges resulting from issues with the theoretical basis of evaluations, for example that foundation models' general capabilities mean that a diverse set of assessments is required to properly assess their performance, capabilities and limitations. Secondly, challenges resulting from practical implementation and engineering issues, for example that the supply chain and the development and deployment lifecycle of foundation models offer several intervention points where evaluations can take place. Thirdly, there are social and policy challenges. For example, the results from a particular set of tests do not guarantee the same behaviour in real-world conditions, and do not predict what will happen if a foundation model is modified. See the table 'Challenges for evaluations' and 'What is foundation model evaluation?' for more detail.

There is not yet consensus on precisely what the term 'evaluation' entails. Evaluation, like 'audit', is a contested term that is defined and understood differently by individuals and communities. There is a narrow view that focuses on tests of the model itself and its outputs, behaviours or alignment with benchmarks, and a broader view that includes tests of a model in a deployed environment to study real-world impacts on users or society.

There is also not yet consensus on a standard set of methods for evaluating foundation models. Current evaluations present models with a variety of inputs and check that the corresponding outputs meet ethical and safety goals. These goals are typically specified by the evaluator or model developer. There are several narrow, specific tests for assessing risks of foundation models, most of which were developed for research purposes and only a few of which were intended for assessing real-world behaviour of a model. Many evaluations seek to 'benchmark' model performance using a set of standardised questions, whereas other evaluations assess a risk through adversarial 'red teaming'.

Benchmarking is a score or metric derived from testing a model on a specific dataset or set of datasets allowing comparisons with other models. Red teaming involves individuals or groups (the ‘red teams’) being tasked with ‘attacking’ a system to find vulnerabilities and flaws. Evaluators can use a mix of approaches to evaluate the same model or target, for example, benchmarking and red teaming can both be used to evaluate bias in foundation models.

In this paper, we take a broad view of evaluations that encompasses their use throughout different stages of the development and deployment lifecycle of a foundation model. We also consider evaluations that look at a wide range of risks and capabilities. This allows us to review several ways that evaluations are conceptualised and used, and to cover a broader range of ways policymakers and regulators can use and engage with evaluations.

Evaluations might be model-focused, exploring the capabilities of a model. While ‘capabilities’ is a poorly defined term in the literature, it broadly refers to the kinds of behaviours, tasks or actions a model is capable of achieving or exhibiting. These kinds of evaluation can assess the absolute capabilities of a model (a model’s behaviour under any conditions), its contextual capabilities (a model’s behaviour under specific conditions) and the propensity of a model to exhibit certain behaviours.

Evaluations might also be broader in focus, ‘zooming out’ from the behaviour of the model itself, assessing how users interact with a foundation model or foundation model application. This could involve reviewing, for example, how much users over-rely on a model’s outputs; how the applications of the model have broader systemic impacts (for instance, the impact on labour markets); or the energy demands of training and using the model.

This paper aims to assess the landscape of foundation model evaluations, across different targets of and approaches to evaluation. It examines the practical, theoretical and social limitations of current approaches. It then aims to assess how evaluations are currently being used by companies and if, and how, they could be a tool for policymakers and regulators. Finally, the paper discusses some possible options for improving evaluations as a tool for AI governance.

Research and findings

Given their growing prominence, we sought to understand how effective and reliable evaluation methods are for ensuring AI systems are safe. Between January and April 2024, we spoke to 16 experts from foundation model providers, third-party evaluation companies, academic labs and civil society organisations, alongside reviewing the foundation model evaluations literature. We learned that:

- Current evaluation methods are a useful tool for better understanding foundation models, but the field has several theory, implementation and social challenges that governments, companies and researchers need to tackle together, to make evaluations a more effective part of the AI governance toolkit (see the table 'Challenges for evaluations' below).

- Evaluations alone are not sufficient for determining the safety of foundation models, the systems built from them and their applications, for people and society in real-world conditions. To be effective, they need to be used alongside other governance tools including codes of practice, incident reporting and post-market monitoring.

- Existing evaluation methods like red teaming and benchmarking have technical and practical limitations. They risk being manipulated or 'gamed' by developers who may train models on the same evaluation dataset that will be used to assess the model (equivalent to seeing the exam paper before the exam); or by strategically choosing which evaluations to use to assess the model.

- It also matters which version of a model is being evaluated. Small changes to an AI application built on a foundation model – including a downstream user fine-tuning the model – can cause unpredictable changes in its behaviour and may override built-in safety features.

- Foundation model safety cannot be tested in a vacuum. Assessing the safety of a model requires considering the wider context, including the users, the design of its interface, what tools the model might have access to, or how the model will affect the environment it operates within. There are valuable tests to be done in a lab setting, and there are important safety interventions to be made at the model level, but they don't provide the full story and need to be paired with context-specific evaluations.

- Current evaluations appear to be designed to meet corporate needs or academic curiosity rather than public or regulatory interests. There are also serious concerns around model transparency by developers, which reduce the ability of third-party assessors to meaningfully evaluate these models.

Challenges for evaluations

Theory	Implementation	Social and policy
Evaluations are not related to real-world applications and harms	Resource intensive (for example, time, cost, labour)	Little involvement of affected communities in designing and conducting evaluations
Difficult to predict relevant risks and harms of general-purpose models	Fine-tuning models can override safety mechanisms[2]	Difficult for evaluators to obtain model or data access
Gaps in evaluation landscape (for example, different modalities, systemic risks, cross-cultural context)	Prompt sensitivity of models makes evaluations less robust	Hard to interpret evaluation metrics/ results and translate to actions
		Lack of incentives to develop evaluations connected to public or regulator interests
		Easy to manipulate evaluations due to lack of transparency

Limitations of benchmarking and red teaming evaluations

Benchmarking	Red teaming
Do not map onto real-world harms in deployment	Difficulty in recruiting red teams with relevant expertise
Lack of robust results due to model outputs' sensitivity to changes in prompts	Expensive and time consuming
Too static to assess interactive dialogues through chatbots	Risks to evaluators' wellbeing
Results may not be trustworthy due to data contamination (for example, models trained on benchmark data)	Difficulty of anticipating risks and methods to exploit a system
Choice of key benchmarks is arbitrary	Lack of diversity of red teams
Not sufficient to meaningfully measure capabilities	Lack of standards and methods to generalise and compare results

[2] Fine-tuning trains a pre-trained model with an additional specialised dataset, removing the need to train from scratch.

How can policymakers and regulators use evaluations?

The goal of emerging global AI governance regimes is to ensure advanced AI systems are safe, effective and lawful. Current evaluation methods are not enough to meet this standard alone, but we recommend that governments, companies and researchers invest in developing evaluations. We are confident that many of the current challenges could be overcome with sufficient effort and rigour.

To improve evaluation methods, it will be necessary to develop context-specific evaluations of AI systems that respond to the needs of specific regulators. This would allow regulators to investigate and assess the safety of particular foundation model applications more effectively. It will also require investment in the underlying science of evaluations, to develop more robust and repeatable evaluations that are based on an understanding of how the foundation model operates, rather than just observations of inputs and outputs.

Finally, the results of an evaluation need to produce meaningful changes in development or deployment, e.g. not releasing a model, or adding safety features. Our interview subjects reported that current evaluations may not always lead to meaningful changes in company practice or release decisions. This can only be addressed through regulation that creates an incentive for companies to take evaluations seriously. National regulators will require new powers to undertake effective independent scrutiny of foundation models and their applications, along with the possibility of blocking the release of models or applications that appear unsafe.

Interviewees for this project noted ways that current evaluations could potentially be used by policymakers and regulators. In descending order of our view of their current viability, they are:

1. As an exploratory tool, for example, to gather evidence for broader risk prioritisation or policymaking.
2. As an investigative tool for regulators scrutinising a particular model or organisation.
3. As part of a licensing or mandatory safety testing regime before a model is made available to the public or sold.

Next steps for regulators and policymakers

While we recognise there are challenges for the current use of evaluations, our research shows that evaluations are a valuable part of the AI governance toolkit that are worth developing further. Below are steps we propose policymakers and regulators should take to make evaluations a more effective tool for scrutinising foundation models and their impacts:

- To help ensure evaluations meet their needs, regulators and policymakers must clearly articulate the insights they seek from evaluations. Simultaneously, the evaluation community must maintain transparency regarding existing limitations and the potential for future advancements.

- To limit the risk of evaluation gaming, policymakers and regulators could keep the details of some evaluation and related datasets confidential. If governments did seek to develop their own evaluation datasets, this would require sufficient in-house expertise and resources.

- Evaluations often fail to reflect the perspectives and experiences of those affected by AI systems. One solution could be for policymakers and regulators to mandate more public participation in the creation of evaluations and the consequences of those evaluation results.

- Funding bodies need to support researchers undertaking fundamental research into evaluation science, including mechanistic and theory-grounded evaluations, methodologies less sensitive to variations in model inputs and wider suites of applications-specific and multi-modal evaluations.

- Governments should implement measures to support an ecosystem of third-party evaluations, including certification schemes and initiatives to ensure assessors have the necessary access to the model, dataset and organisational information to conduct an evaluation.

How to read this report

This report provides insights from interviews with experts and a review of literature on foundation model evaluations. This is a long document. Depending on your background and interests, we recommend different reading strategies:

For all readers (10–15 minute read):

- ‘Executive summary’ for key findings. This provides a concise overview of the entire report’s main points and conclusions.
- ‘What are foundation models?’ for context. This section helps establish a common understanding of the technology being discussed.
- ‘What is foundation model evaluation?’ for an overview. This gives you grounding knowledge of the evaluation process, which is crucial for understanding the rest of the report.

If you are a policymakers or regulator…
…and you’re new to foundation model evaluations (30–45 minutes):

- ‘Executive summary’: This gives you a high-level view of the key issues and findings.
- ‘What is foundation model evaluation?’ This chapter provides essential background knowledge for understanding the rest of the report.
- ‘What are the most common approaches to evaluation?’ This helps you understand the methodologies currently in use, which is crucial for informed policymaking.
- ‘What do current evaluations aim to assess?’ This section outlines the current focus areas of evaluations, helping you identify potential gaps in coverage.

…and you’re considering using evaluations in your work or relying on their results as part of your policy work (60–90 minutes):

- Start with ‘What are the challenges for current evaluations?’, then skim specific challenges relevant to your work. Understanding these challenges is crucial for developing robust policies and avoiding potential pitfalls.

- 'What is the role of evaluations in the broader landscape of AI governance and accountability?' This chapter places evaluations in the wider context of AI governance, helping you understand their potential impact.
- 'How should regulators and policymakers think about using evaluations?' This provides suggestions on how evaluations can be incorporated into regulatory frameworks.
- 'Making evaluations a more effective part of the governance toolkit'. This chapter offers practical suggestions for improving the effectiveness of evaluations in governance. Read at least 'Asking questions', 'Involving affected communities' and 'Disclosure requirements and external scrutiny of company claims' for suggestions about how your use of evaluations can start from a higher bar.

If you are an AI researcher or developer (30–45 minutes):

- 'What are the challenges for current evaluations?' This section summarises key insights from our interviews and literature review on the limitations and difficulties faced in current evaluation practices.
- 'What is the role of evaluations in the broader landscape of AI governance and accountability?' This provides context on how evaluations are being used and perceived beyond the research community.
- 'Building a science of evaluations'. This section reflects on the need for more fundamental research in evaluation methodologies, as highlighted by our interviewees.
- 'Making evaluations a more effective part of the governance toolkit'. While primarily aimed at policymakers, this chapter includes recommendations that may influence future research directions and collaborations with regulators.

If you are an AI firm executive or decision-maker (30–45 minutes):

- 'Company actions as a result of evaluations'. This section provides insights from our interviews on how companies are using evaluations in their decision-making processes.
- 'Appendix 1: Structured approaches to development and deployment decisions based on evaluations'. This reflects on current industry practices for integrating evaluation results into decision-making.
- 'What are the challenges for current evaluations?' Understanding these challenges, as reported by our interviewees, is crucial for interpreting and using evaluation results effectively.

- 'Making evaluations a more effective part of the governance toolkit'. While aimed at policymakers, this chapter includes recommendations that may affect future regulatory expectations for AI companies.

If you are third-party evaluators (considering) working on foundation model evaluations (30–45 minutes):

- 'Who is involved in evaluations?' This provides an overview of the different roles in the evaluation ecosystem based on our research, helping you understand how your role is perceived.
- 'What are the challenges for current evaluations?' This section summarises key insights from our interviews on the difficulties faced in conducting effective evaluations.
- 'Building up the third-party ecosystem'. This section discusses the potential future of third-party evaluations based on our interviews and literature review.
- 'Making evaluations a more effective part of the governance toolkit'. While primarily aimed at policymakers, this chapter includes recommendations that may shape the future demand for third-party evaluation services.

Introduction

> ‘Safety testing and evaluation of advanced AI is a nascent science, with virtually no established standards of best practice. AISI’s evaluations are thus not comprehensive assessments of an AI system’s safety, and the goal is not to designate any system as “safe”.’
>
> The UK’s AI Safety Institute[3]

Whether a smartphone, a prescription drug or a car, we expect the products we use to be safe and reliable. The Ada Lovelace Institute’s public attitudes research shows the UK public expect AI-powered applications and services that impact our everyday lives to be explainable, contestable and subject to independent oversight.[4] People expect data-driven innovation to be ethical, responsible and focused on public benefit.[5]

However, evaluating AI systems for this purpose is not an easy task. An AI system may pose different kinds of risks that depend on the context in which it is deployed. Different risks can arise at different stages of an AI system’s development and use. Unlike for example, paracetamol, where millions of uniform tablets might roll off a production line with intended uses, defined risks and benefits, each deployment of an AI system may be different, meaning that an issue identified in one version of an AI system may not exist in another. This is particularly true when downstream developers ‘fine-tune’ AI foundation models to build custom AI applications and services.

[3] AI Safety Institute, ‘AI Safety Institute Approach to Evaluations’ (GOV.UK, 9 February 2024) https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations accessed 1 March 2024.

[4] Ada Lovelace Institute and Alan Turing Institute, ‘How Do People Feel about AI? A Nationally Representative Survey of Public Attitudes to Artificial Intelligence in Britain’ (2023) 39–44 https://www.adalovelaceinstitute.org/report/public-attitudes-ai/ accessed 6 June 2023.

[5] Ada Lovelace Institute, ‘Who Cares What the Public Think?’ (2022) https://www.adalovelaceinstitute.org/evidence-review/public-attitudes-data-regulation/

Foundation models are capable of a range of general tasks such as text synthesis, image manipulation and audio generation.[6] Notable examples are GPT-3.5 and GPT-4, OpenAI's families of foundation models that underpin the conversational chat agent ChatGPT, and many other applications via OpenAI's application programming interface (API).

Companies building foundation models may offer a whole range of models which are optimised for different levels of speed, cost and performance. For example, Google's 'most capable' family of foundation models, Gemini 1.0, offers Nano, Pro and Ultra versions that are tailored for different use cases by downstream developers via its Google Cloud API. These models are frequently tweaked or changed with a range of safety filters. Different versions of this model can also be fine-tuned for specific tasks by users via an API. Other companies have released their foundation models via open source-like licences. For example, Meta has released its LLaMa-3 foundation model via an open source-like licence, making it available via model hosting services such as HuggingFace and GitHub.

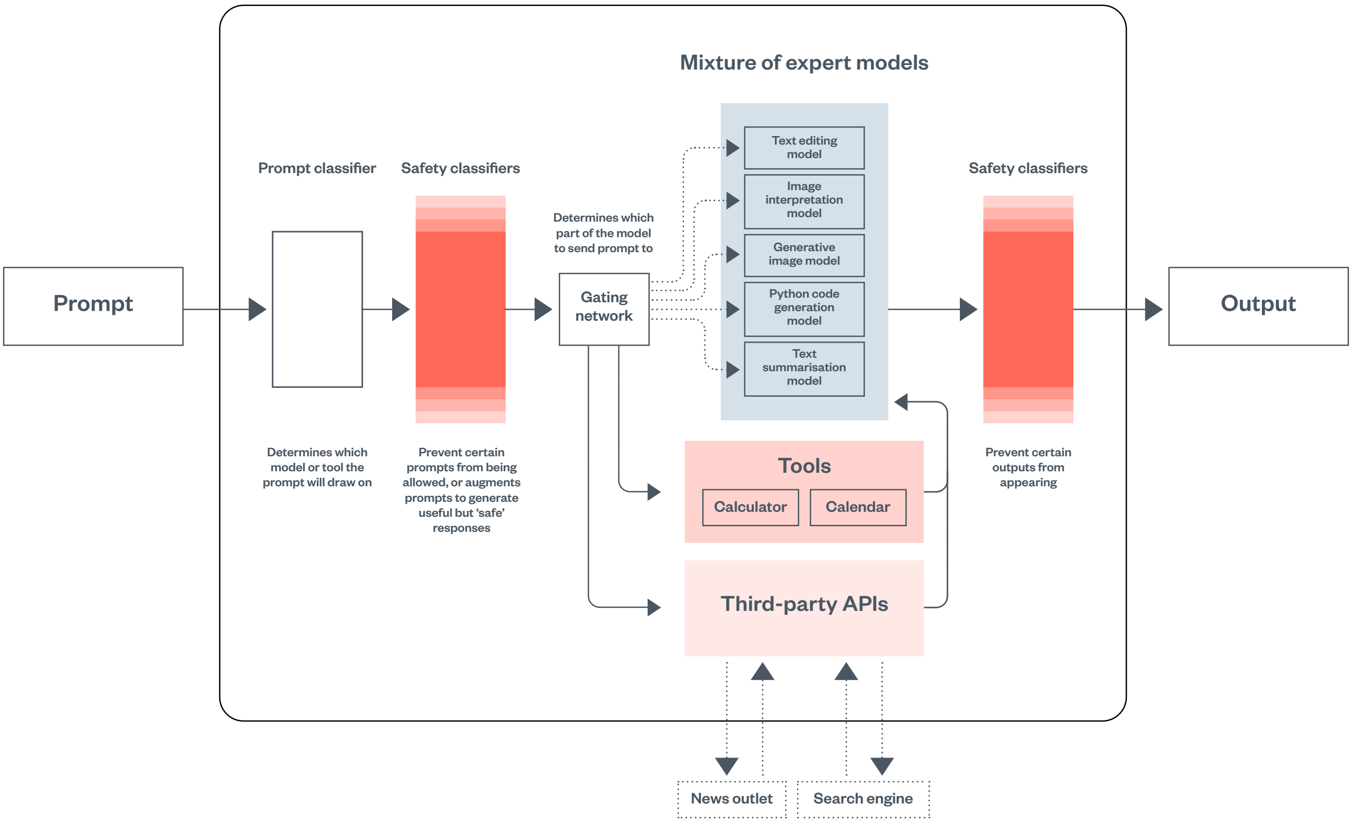
> It is important to note that the version of a foundation model made accessible to users and developers is often a filtered model that includes additional components.

These components can include prompt classifiers, safety classifiers (which prevent certain inputs from reaching the model or certain model outputs reaching the end-user) and tools like calculators, calendars or even access to other APIs from news outlets and search engines. These components enable or disable certain kinds of interactions with the underlying model. This is a crucial aspect when considering foundation model evaluations, as the additional components may cause changes in model behaviour or change the distribution of outputs for a given set of inputs, when compared with an unfiltered model.

[6] To learn more about foundation models, you can read our explainer on: "What is a foundation model".

This is illustrated in Figure 1 below, which shows the stages an inputted prompt might go through in a foundation model before an output is shared with the user. The figure below uses a 'Mixture of Experts' model, where the foundation model is divided into sub-models (each specialised for a particular set of tasks) and inputs are routed the most relevant sub-model. Safety classifiers prevent certain prompts or outputs that are deemed unsafe.

Figure 1: Foundation models at a glance


Mixture of expert models
Prompt classifier
Safety classifiers
Text editing model
Image interpretation model
Generative image model
Python code generation model
Text summarisation model
Safety classifiers
Determines which part of the model to send prompt to
Prompt
Gating network
Output
Determines which model or tool the prompt will draw on
Prevent certain prompts from being allowed, or augments prompts to generate useful but 'safe' responses
Tools
Calculator
Calendar
Prevent certain outputs from appearing
Third-party APIs
News outlet
Search engine

As foundation models are capable of a wide range of general tasks, they differ from narrow AI systems that focus on a specific or limited task. These include older predictive text or image recognition models. Some key differences between foundation models and narrow AI systems include:

- It can be harder to identify and foresee the ways foundation models can benefit people and society.
- It is harder to predict how foundation models may be used, in what contexts they will be deployed, how they can affect end users and other people impacted by the system, and therefore when they can cause harm.

- Foundation models' failures could have systematic, cascading effects, if hundreds of applications depend on a single foundation model.

Nevertheless, policymakers are incorporating evaluations into emerging regulatory proposals in the EU, UK and USA. These jurisdictions are creating voluntary and legally mandated requirements for developers to evaluate AI systems for different kinds of risks.

The EU AI Act includes requirements for evaluation and testing of datasets and models that are high risk. There are similar proposals in the USA and the UK. Both countries have secured voluntary commitments from major AI companies to allow external evaluations of the companies' AI models. The UK, USA, Japan and Singapore have also launched separate AI safety institutes with a mission to develop novel methods for testing advanced AI systems for different kinds of risks, with more institutes to follow from France and Canada.

But how effective are evaluation methods at assessing and mitigating the risks of foundation models? As the quote above from the UK's AI Safety Institute suggests, and as interviewees from leading companies and independent evaluators recognised, the field of foundation model evaluation is still nascent and maturing.

> Current evaluations alone are not up to the task of guaranteeing or even effectively testing whether a model is safe to release.

However, there is exploratory work underway from AI companies, independent researchers and organisations like the UK's AI Safety Institute to make evaluations more scientifically grounded and rigorous.[7]

What 'safety' means in the context of AI is also contested and the term safety is used differently by different groups. AI-based systems safety literature refers to preventing a system from causing harm to humans,

[7] Will Henshall, 'Nobody Knows How to Safety-Test AI' (TIME, 21 March 2024) https://time.com/6958868/artificial-intelligence-safety-evaluations-risks/ accessed 22 March 2024.

the environment or monetary assets.8 In this safety literature, AI systems are sociotechnical systems with many components, including the model itself, the technical scaffolding and tools around the models, the humans who interact with these systems, the laws and norms that govern its use, and many other components beyond the purely technical aspects. For this reason, approaches to safety should seek to address all aspects of these components of AI systems.

It is crucial to understand the limits and opportunities of evaluations, and what other mechanisms may be needed to create a more holistic approach to assuring the safety of AI systems.

In this paper, we sought to answer five core questions:

1. Compared to 'narrow' AI systems, what are the similar and unique risks that foundation models pose for people and society?
2. What is the difference between evaluation and other forms of assessment and accountability?
3. What are the proposed range of evaluation and testing approaches for addressing the risks of foundation models?
4. What are the limitations of proposed approaches to evaluation and testing?
5. What measures can be taken by policymakers to create legal/regulatory accountability of different actors in the foundation model lifecycle based on the results of evaluation and testing?

We spoke with 16 experts from academic, industry and civil society, and conducted a literature review on the state of evaluations (see the 'Methodology' section for more details and our participant information table). This paper addresses these questions and concludes with some recommendations for policymakers on how to advance the state of the science of evaluations.

[8] Heidy Khlaaf, 'Toward Comprehensive Risk Assessments and Assurance of AI-Based Systems' (Trail of Bits 2023) https://www.trailofbits.com/documents/Toward_comprehensive_risk_assessments.pdf

What is foundation model evaluation?

What is and isn't an evaluation?

As large-scale foundation models are a relatively new development, it is unsurprising that the evaluation of foundation models is a nascent field. Foundation models are created using enormous amounts of training data, compute resources and complex algorithmic architectures. They can exhibit complex, unpredictable and contradictory behaviour when prompted by human users.[9] They are trained on internet-scale datasets containing billions or trillions of words that are too large for any individual to thoroughly understand.[10] Slight changes in inputs can result in significant changes in outputs.[11] All of these make foundation model behaviour and interaction challenging to evaluate.

Our interviews and literature review found that there is not yet a consensus on precisely what the term 'evaluation' entails. Some of our interviewees took a narrow view of evaluations that focused on tests of the model itself and its outputs, behaviours or alignment with benchmarks.[12] Others took a broader view that included tests of a model in a deployed environment to study its real-world impacts on users or society.[13]

There was also no consensus on a standard set of methods for evaluating foundation models.[14] Many evaluations seek to benchmark

[9] Milad Nasr and others, 'Scalable Extraction of Training Data from (Production) Language Models' (arXiv, 28 November 2023) http://arxiv.org/abs/2311.17035 accessed 27 June 2024.

[10] Jesse Dodge and others, 'Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus', *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics 2021) https://aclanthology.org/2021.emnlp-main.98 accessed 22 March 2024.

[11] Pengfei Liu and others, 'Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing' (2023) 55 ACM Computing Surveys 195:1.

[12] P6, P5, P9

[13] P15, P2

[14] Neel Guha and others, 'AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing' (15 November 2023) https://papers.ssrn.com/abstract=4634443 accessed 22 March 2024.

model performance using a set of standardised questions, whereas other evaluations assess a risk through adversarial red teaming. There are several narrow, specific tests for assessing risks of foundation models, most of which were developed for research purposes and only a few of which were intended for assessing real-world behaviour of a model.

Even so, there were areas of agreement among our interview participants. Most interviewees agreed that evaluation involves some form of, ideally systematic, assessment of a model.[15] Whether narrow testing or broader analysis, there was agreement that the core goal is to understand the foundation model and/or its impacts, including the model's capabilities, risks, performance, behaviour and social impact. Interviewees also agreed that no single evaluation method gives this whole picture.[16]

Interviewees also agreed that evaluations do not happen in a vacuum and that many factors impact evaluations. The interests and motivations of those conducting the evaluation shape the choice of goals, methodology used and what is in scope.[17] For example, the primary goal of some evaluations may be to demonstrate comparative effectiveness of the model versus commercial competitors (for example, performance on particular public benchmarks). Other evaluations seek to identify potential misuse of the model by malign actors, such as whether the use of a model to plan biological weapon attacks substantially increases the risk of an attack, compared with only using the internet.[18]

How are evaluations different from audits?

While foundation model evaluations are a nascent field, they can often be confused with algorithmic audits, another method for testing AI systems for certain risks. Both 'evaluation' and 'audit' are contested terms that are defined and understood differently by individuals and communities. Some interviewees used the terms interchangeably or used the term

[15] P4, P9, P14, P12

[16] P4, P14, P16, P7

[17] P14, P15, P1, P2

[18] Christopher A Mouton, Caleb Lucas and Ella Guest, 'The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study' (RAND Corporation 2024) https://www.rand.org/pubs/research_reports/RRA2977-2.html accessed 20 March 2024.

Audits are seen as broader than evaluations

audit more than evaluation in their answers.[19] However most interviewees identified differences between the two.[20]

Our interviewees acknowledged the way these terms are used in the context of AI can be vague or confusing;[21] whereas in existing regulated industries, for example, finance, audit is generally understood to refer to a specific and well-defined process.[22] Interviewees expressed varying views about the relationship between audits and evaluations, highlighting the lack of consensus in this field. Two interviewees described evaluations as a subset or component of the auditing process,[23] while two different interviewees saw audits as a subset of evaluations.[24] Some interviewees described audits as more structured than evaluations with a clearer understanding of what auditors were looking for.[25] On the other hand, one interviewee thought of audits as more dynamic and flexible in their approach compared to more standardised evaluations.[26]

There was more consensus on targets of audits, which were described as broader than evaluations. The clearest distinction was that evaluations related to properties of the model itself: its behaviour, or the consequences of its inputs and outputs. Whereas audits could also include assessment of the development and deployment processes and governance practices.[27] For example, the audits that were discussed included: compliance audits, when a model is assessed for its compliance with regulatory requirements; conformity assessment, when a system is assessed for whether it fulfils certain standards; and security audits, when a system is checked for security vulnerabilities.[28]

[19] P1, P5, P13
[20] P3, P4, P6, P7, P10, P12, P14, P15, P16
[21] P1, P5, P14
[22] P1, P5
[23] P6, P15
[24] P4, P16
[25] P3, P7
[26] P10
[27] P7, P10, P12, P14
[28] P1, P3, P5, P15

> In this paper, we take a broad view of evaluations that encompasses their use throughout different stages of the development and deployment lifecycle of a foundation model.

We also consider evaluations that look at a wide range of risks and capabilities. This allows us to review several ways that evaluations are conceptualised and used, and to cover a broader range of ways policymakers and regulators can use and engage with evaluations.

When does the evaluation take place? What is being evaluated?

Foundation model evaluations can happen across the supply chain of a foundation model, and at different deployment and development stages. Some approaches to evaluation take a purely technical approach, focusing on the inputs of the AI model, such as bias in the training data; or the outputs of the AI system, such as the text output by a language model in response to a prompt.[29]

Other approaches aim to evaluate a system in context, to understand how an AI system or AI application impacts individual users. Some approaches go further and aim to understand the broader societal impacts of the AI system. For example, an evaluation could aim to assess the net environmental impacts of an AI system's development and deployment. This could include measurement of direct energy and water usage, combined with an estimation of counterfactual usage of alternatives to the AI system.

Foundation model evaluations can assess multiple 'components' throughout the AI development process. These include:

- training data – for example, to assess potential biases, representativeness, presence of harmful content
- base models – for example, to assess core capabilities, tendencies towards generating unsafe outputs, vulnerability to manipulation

[29] Laura Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (arXiv, 31 October 2023) 7–8 http://arxiv.org/abs/2310.11986 accessed 8 November 2023.

- fine-tuned models – for example, to assess whether new risks emerge during task adaptation or if safety mechanisms remain effective after additional training
- systems and applications – for example, to assess real-world usability, security, fairness and broader social impacts.[30]

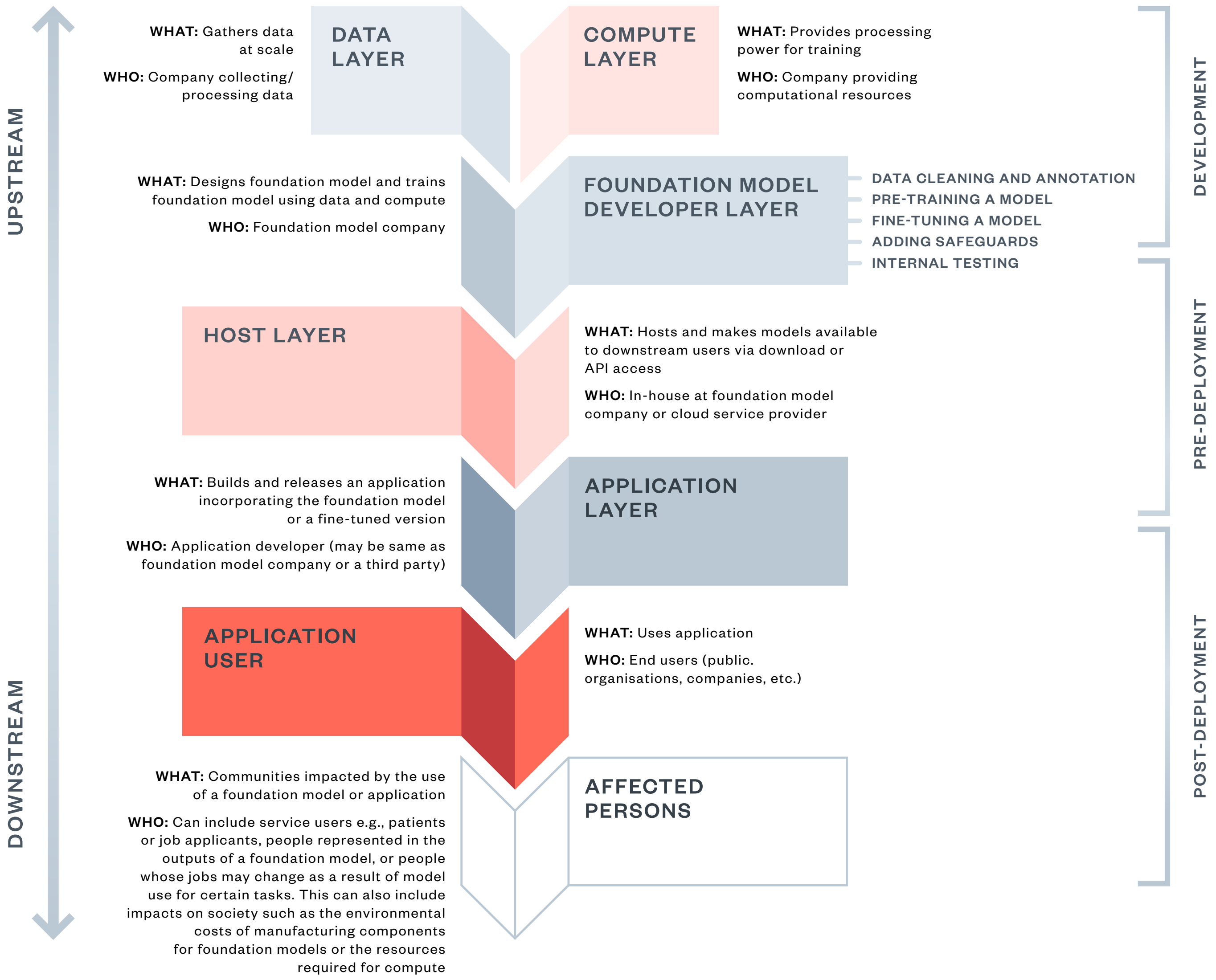
These can be assessed at different stages of the foundation model supply chain. There are multiple stages in the foundation model supply chain, from the collection and aggregation of data and acquisition of compute, through to development and deployment to the end-users and those affected by the outputs of the foundation models.[31]

[30] AI systems include not only the weights and architecture of the model, but also include a broader set of system parameters. These can include retrieval databases and particular kinds of prompts. Lee Sharkey and others, 'A Causal Framework for AI Regulation and Auditing' (Apollo Research 2023) 4 https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/65a6f1389754fc06cb9a7a14/1705439547455/auditing_framework_web.pdf

[31] Elliot Jones, 'Explainer: What Is a Foundation Model?' (*Ada Lovelace Institute*, 17 July 2023) https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/ accessed 1 August 2023.

Figure 2: The foundation model supply chain

Note: This is one possible model (there will not always be a separate or single company at each layer)


UPSTREAM
DOWNSTREAM
DATA LAYER
WHAT: Gathers data at scale
WHO: Company collecting/processing data
COMPUTE LAYER
WHAT: Provides processing power for training
WHO: Company providing computational resources
FOUNDATION MODEL DEVELOPER LAYER
WHAT: Designs foundation model and trains foundation model using data and compute
WHO: Foundation model company
DATA CLEANING AND ANNOTATION
PRE-TRAINING A MODEL
FINE-TUNING A MODEL
ADDING SAFEGUARDS
INTERNAL TESTING
HOST LAYER
WHAT: Hosts and makes models available to downstream users via download or API access
WHO: In-house at foundation model company or cloud service provider
APPLICATION LAYER
WHAT: Builds and releases an application incorporating the foundation model or a fine-tuned version
WHO: Application developer (may be same as foundation model company or a third party)
APPLICATION USER
WHAT: Uses application
WHO: End users (public. organisations, companies, etc.)
AFFECTED PERSONS
WHAT: Communities impacted by the use of a foundation model or application
WHO: Can include service users e.g., patients or job applicants, people represented in the outputs of a foundation model, or people whose jobs may change as a result of model use for certain tasks. This can also include impacts on society such as the environmental costs of manufacturing components for foundation models or the resources required for compute
DEVELOPMENT
PRE-DEPLOYMENT
POST-DEPLOYMENT

This supply chain and the development and deployment life cycle of foundation models offer several intervention points where evaluations can take place:[32]

[32] Markus Anderljung and others, 'Towards Publicly Accountable Frontier LLMs: Building an External Scrutiny Ecosystem under the ASPIRE Framework' (arXiv, 15 November 2023) http://arxiv.org/abs/2311.14711 accessed 30 November 2023.

- Before training: Evaluating data issues and potential social impacts of forecasted capabilities of the models and its applications. These might happen at the data layer or the foundation mode developer layer.
- During training: Looking at model checkpoints. These mainly happen at the foundation model developer layer.
- Pre-deployment: Models and AI systems can be tested for potential risks, harms and capabilities. Testing can happen at the foundation model developer layer or host layer.
- Post-deployment: Effective assessment of real-world impact necessitates ongoing monitoring. Evaluations might be re-run at fixed time intervals, for example, when the model is fine-tuned or when new system features are added. These would happen at the application layer or by the application user, sometimes in collaboration with the foundation model developer.

Understanding these intervention points could be important for policymakers, as potential regulations could target specific stages with differing evaluation requirements.

What does it mean to evaluate ‘capabilities’?

Model-focused evaluations are a valuable tool for exploring the capabilities of a model. While ‘capabilities’ is a poorly defined term in the literature, it broadly refers to what kinds of behaviours, tasks or actions a model is capable of achieving or exhibiting. Evaluations can try to assess the absolute capabilities of a model (a model’s behaviour under any conditions), its contextual capabilities (a model’s behaviour under its existing conditions) and/or the propensity of a model to exhibit certain behaviours.

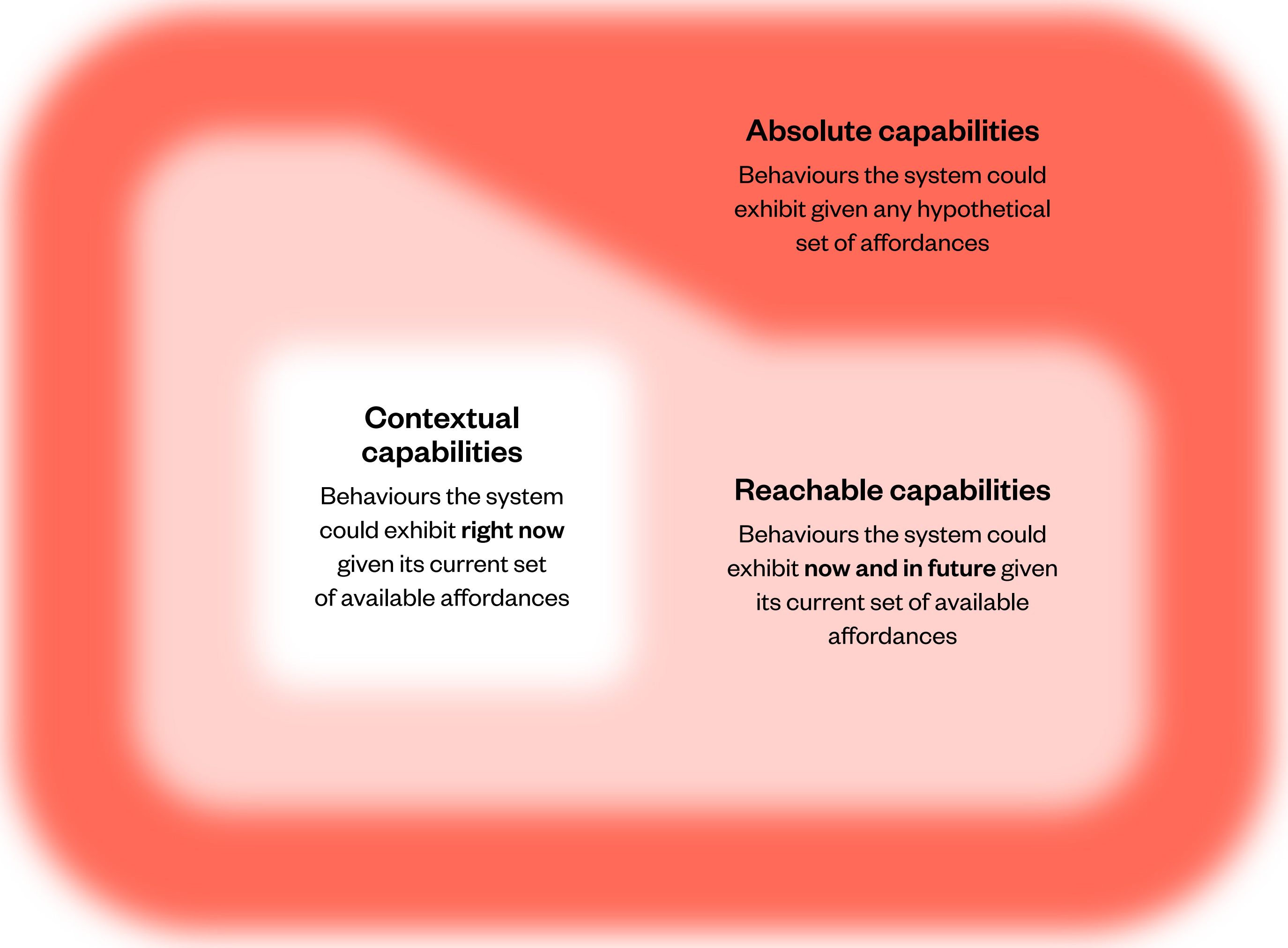
In this paper, we adopt Sharkey and others’ (2023) framework for conceptualising AI system capabilities. In this framework, they define AI system behaviour as the set of actions or outputs that a system actually produces and the context in which they occur; for example, an LLM outputting a particular sentence in response to a particular input prompt. They define AI system affordances as the environmental resources and opportunities for influencing the world that are available to a system; for example, the design of interface, the guardrails that restrict inputs and outputs, and access to plugins that allow it to use a calculator or search the web.

Sharkey and others (2023) then distinguish between the absolute capabilities, reachable capabilities and contextual capabilities of an AI system.33 Figure 3 below shows how absolute capabilities covers the whole set of potential behaviours an AI system could exhibit, given any and all affordances, regardless of whether those affordances are currently available. Reachable capabilities are then a subset of absolute capabilities: the behaviours the AI system could exhibit now or in the future given its current available affordances and current environment. Contextual capabilities are then finally a subset of reachable capabilities, the behaviours the AI system can exhibit right now given its current set of affordances in its current environment.

Finally, the AI system propensities are then the tendency of a system to express one behaviour over another, see Figure 4. Even though systems may be capable of a wide range of behaviours, they may in fact be biased towards certain kinds of behaviour. For example, a foundation model such as GPT-4 or Claude 3 might be in theory capable of producing discriminatory content; however, through fine-tuning, the model may be trained to almost always refuse to produce discriminatory content when prompted to do so.

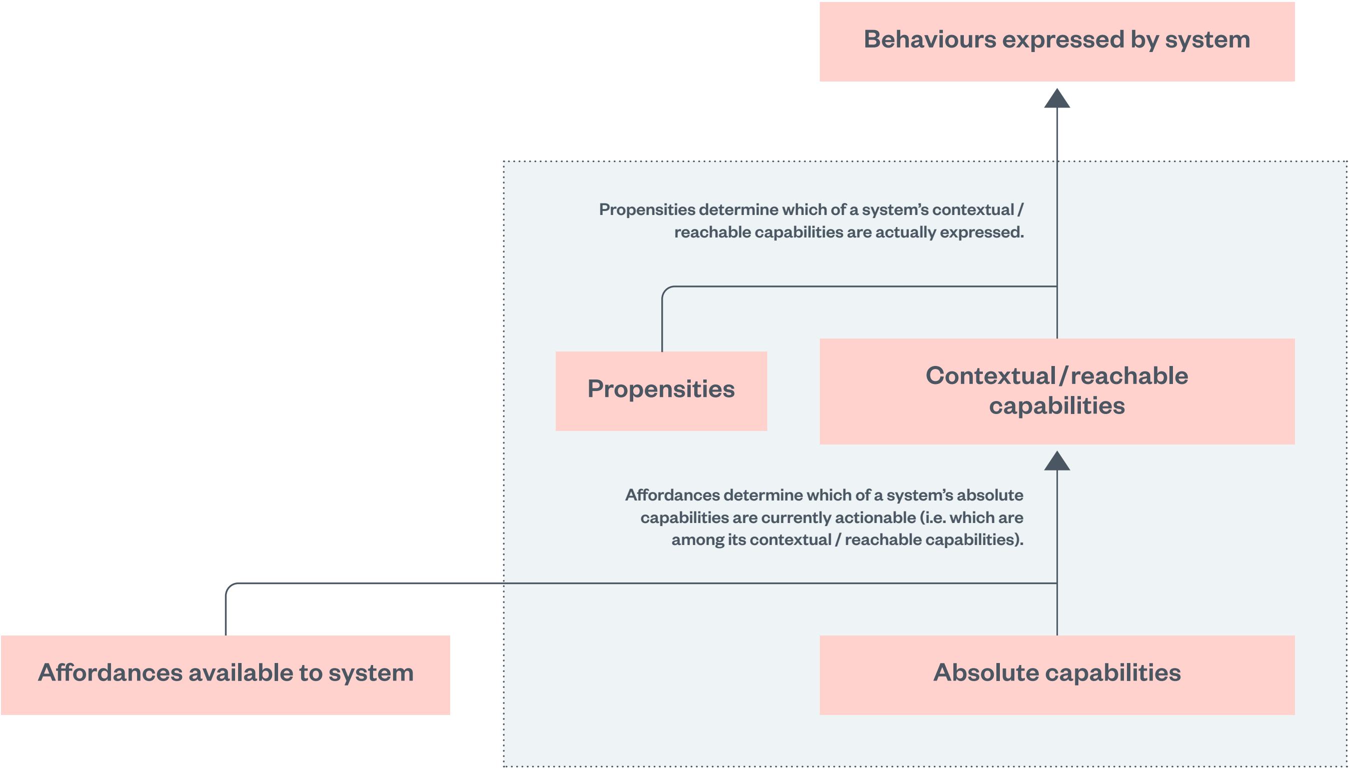
[33] Sharkey and others (n 31) 5.

Figure 3: The relationship between the sets of potential behaviours defined by absolute capabilities, reachable capabilities and contextual capabilities – reproduced from Sharkey and others (2023)[34]


Absolute capabilities
Behaviours the system could exhibit given any hypothetical set of affordances
Contextual capabilities
Behaviours the system could exhibit **right now** given its current set of available affordances
Reachable capabilities
Behaviours the system could exhibit **now and in future** given its current set of available affordances

[34] ibid 6.

Figure 4: The relationship between an AI system's capabilities, propensities, affordances and behaviours – reproduced from Sharkey and others (2023)[35]


Behaviours expressed by system
Propensities determine which of a system's contextual / reachable capabilities are actually expressed.
Propensities
Contextual / reachable capabilities
Affordances determine which of a system's absolute capabilities are currently actionable (i.e. which are among its contextual / reachable capabilities).
Affordances available to system
Absolute capabilities

Evaluations can involve capability elicitation, where evaluators add post-training enhancements to the model to achieve more capable performance on a given task, giving a better sense of the model's overall capabilities. This can involve prompt engineering and optimisation, fine-tuning the model or providing additional affordances to the model.[36]

This process generally aims to measure what capabilities might be reachable with moderate amounts of effort and/or additional affordances, not necessarily the absolute capabilities of the model.[37]

[35] ibid.

[36] ibid 11; 'Guidelines for Capability Elicitation' (*METR's Autonomy Evaluation Resources*, 15 March 2024) https://metr.github.io/autonomy-evals-guide/elicitation-protocol/ accessed 28 May 2024; Toby Shevlane and others, 'Model Evaluation for Extreme Risks' (arXiv, 22 September 2023) 13 http://arxiv.org/abs/2305.15324 accessed 9 January 2024.

[37] 'Guidelines for Capability Elicitation' (n 37).

Capability elicitation generally does not aim to estimate the propensity of foundation model behaviours.[38] Some researchers have criticised the extrapolation from a model performing a particular task to a model having a capability, calling for more rigorous taxonomy and grounding of claims around capabilities.[39]

Capability elicitation is often used in the context of dangerous capability evaluations, which generally aim to demonstrate that an AI system can exhibit the dangerous capability at all, for example, the ability to generate persuasive phishing emails or produce bioweapon-relevant information.[40]

Who is involved in evaluations?

Evaluators can be categorised as first-, second- and third-party evaluators, each with their own roles and motivations.41 The following taxonomy provides a guide for evaluators, although it is subject to change due to the lack of consensus on terminology in this area.

First-party evaluations

First-party evaluations are run within organisations to evaluate their own applications or services. In an AI context, this means evaluating their own data, their own models and their own applications.

Examples of first-party evaluators

- Model development teams: Data scientists, machine-learning engineers and developers test their own models as part of the training process. Testing could involve conducting performance evaluations for a large language model (LLM): for example, prediction accuracy on text completions at different stages of training; or safety evaluations, for

[38] Sharkey and others (n 31) 11.

[39] Usman Anwar and others, 'Foundational Challenges in Assuring Alignment and Safety of Large Language Models'.

[40] Sharkey and others (n 31) 11.

[41] This section draws heavily on Inioluwa Deborah Raji and others, 'Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance', *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2022) 558 https://doi.org/10.1145/3514094.3534181 accessed 6 October 2023.

example, likelihood of generating toxic content.[42]

- In some organisations, these functions might be carried out by sub-teams and/or by governance functions that are separate to the development team, such as:
  - Performance evaluation teams: Data scientists, machine-learning engineers and developers test models on benchmark datasets during development to optimise performance metrics like accuracy.
  - Internal compliance teams: Groups within an AI development company test models for fairness, transparency, explainability and other ethical AI standards before deployment.

Second-party evaluations

Second-party evaluations are performed by or on behalf of organisations with a commercial interest in the model to be evaluated. This could include a developer providing a customer with access to a model; or a contracted organisation undertaking an evaluation on behalf of a customer. Second-party evaluations tend to be more formal than first-party evaluations because the terms of an evaluation are set out in a contract; and because the evaluation results could influence a customer's purchasing decisions.

Examples of second-party evaluators

- Paid evaluation consultancies: A potential customer hires an external company to evaluate potential harms related to an AI system. The consultancy can act on behalf of a client who is considering purchasing or implementing an AI system.
- Potential purchasers: Government agencies, companies or other groups evaluate supplier AI systems as part of a procurement process to determine fitness for purpose, for example, requiring an assessment of the robustness of an AI system.

[42] Dirk Groeneveld and others, 'OLMo: Accelerating the Science of Language Models' (arXiv, 27 February 2024) 5, 8–10 http://arxiv.org/abs/2402.00838 accessed 30 April 2024.

Third-party evaluations

Third-party evaluations are evaluations by organisations that are external to, and independent of (at least formally) the customer–supplier relationship. Third-party evaluations can fall into one of three categories:

Evaluatee-driven third-party evaluations: In this case, the company whose models or systems are being evaluated hires the evaluator. They might give the third-party evaluators a specific objective (investigate specific cybersecurity-relevant capabilities); ask them to conduct a series of evaluations to certify compliance with an independent standard; or give them a wide remit to conduct exploratory evaluations.

Example of evaluatee-driven third-party evaluator

- Paid evaluation consultancies: External companies hired by the developer company to evaluate potential harms related to an AI system. They operate based on an agreement with the company whose models are being evaluated and that company is the primary audience of the evaluation results.

In second-party evaluations, the evaluation consultant is likely to take a critical or adversarial approach to evaluating the model, testing the limits of the model for the client buying or using the model. In evaluatee-driven third-party evaluations, the degree to which the evaluation consultant is critical will depend on the explicit and implicit instructions of the company whose model is being evaluated. In some cases, they may want a critical and adversarial stress-test of their system, but in other cases they may want simply want results that appear to demonstrate compliance or superior performance, regardless of the rigour of those results.

Government-driven third-party evaluations: In this case, the government or an independent regulator either appoints an evaluator to scrutinise a company's model or systems or directly evaluates the models and systems themselves. In either case, it is the government or an independent regulator choosing what evaluations are conducted.

Example of government-driven third-party evaluator

- Regulators: Government agencies like the US Federal Trade Commission evaluating AI systems to determine regulatory compliance, or investigations into companies or evaluations before granting a licence.

Independent third-party evaluations: In this case, independent organisations (for example, academics, journalists or civil society organisations) undertake evaluations of the models and systems based on their own priorities, with or without the cooperation or knowledge of the developer.

Examples of independent third-party evaluators

- Academic researchers: Scientists testing commercial AI systems independently without being paid by the company, such as those assessing racial bias in healthcare algorithms.
- Investigative journalists: Reporters evaluating AI tools (such as hiring algorithms and content moderation systems) independently to uncover issues.
- Civil society organisation: Organisations performing external evaluations of AI systems' societal impacts.

What are the most common approaches to evaluation?

Foundation models are designed to be general, working across many complex tasks and domains.[43] As a result, evaluating their performance requires a diverse set of assessments to properly assess their capabilities and limitations. However, while other industries like aerospace and medicine rely on strong theoretical underpinnings to prove the generality and validity of safety tests, the theoretical understanding of foundation models is currently lacking.[44] Consequently, results from a particular set of tests do not guarantee the same behaviour in the real-world conditions those tests are designed to imitate, and do not predict what will happen if a foundation model is modified.[45]

Policymakers and AI companies are currently seeking to use evaluations as a method for providing clarity on appropriate and responsible uses of foundation model applications. However, there is no agreed set of methods in the literature for evaluating foundation models.46 Current evaluations of foundation models rely on presenting models with a variety of inputs and checking that the corresponding outputs meet ethical and safety goals. These goals are typically specified by the evaluator or model developer.

There are several different approaches to evaluating foundation models including:

- benchmarking
- red teaming

[43] Rishi Bommasani and others, 'On the Opportunities and Risks of Foundation Models' (arXiv, 12 July 2022) http://arxiv.org/abs/2108.07258 accessed 30 January 2023.

[44] P1, P2, P3, P4, P5, P6, P7, P9, P10, P11, P13, P15

[45] P1, P4, P5, P7, P9, P10, P14, P15, Guha and others (n 10).

[46] Guha and others (n 15).

- bug and bias bounties[47]
- human interaction evaluations[48]
- system-level evaluation approaches to assess the economic, social and environmental impacts of foundation models.[49]

Benchmarking and red teaming are the two most common evaluation approaches identified by interviewees

Some of these approaches are better suited to evaluating how a model is deployed, such as human interaction evaluations, which aim to understand the model's effects on people using and interacting with AI systems. Other approaches such as benchmarking are more focused on evaluating the model in isolation.

In this section we discuss the two most common approaches raised in our interviews, **benchmarking** and **red teaming**. This reflects an evaluation gap where human interaction and system-level evaluations are rare compared to evaluations which centre the model.[50] While foundation model developers sometimes conduct evaluations that aren't centred on the model, most of their public discussion on evaluations is around model-centric evaluations such as benchmarking and red teaming.

Benchmarking is a score or metric derived from testing a model on a specific dataset or set of datasets allowing comparisons with other models. On the other hand, red teaming involves individuals or groups (the 'red teams') being tasked with 'attacking' a system to find vulnerabilities and flaws.

Evaluators can use a mix of approaches to evaluate the same model or target, for example: benchmarking and red teaming can both be used to evaluate bias in foundation models.

[47] 'Keeping GenAI Technologies Secure Is a Shared Responsibility | The Mozilla Blog' https://blog.mozilla.org/en/mozilla/keeping-genai-technologies-secure-is-a-shared-responsibility/ accessed 6 June 2024; 'Humane Intelligence Algorithmic Bias Bounty' (*Humane Intelligence*) https://www.humane-intelligence.org/bounty1 accessed 30 May 2024.

[48] Lujain Ibrahim and others, 'Beyond Static AI Evaluations: Advancing Human Interaction Evaluations for LLM Harms and Risks' (arXiv, 27 May 2024) http://arxiv.org/abs/2405.10632 accessed 31 May 2024; Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (n 30).

[49] Tyna Eloundou and others, 'GPTs Are GPTs: Labor Market Impact Potential of LLMs' (2024) 384 Science 1306; Sabrina Habib and others, 'How Does Generative Artificial Intelligence Impact Student Creativity?' (2024) 34 Journal of Creativity 100072; Alexandra Sasha Luccioni, Sylvain Viguier and Anne-Laure Ligozat, 'Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model'.

[50] Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (n 30).

> A range of evaluations for foundation models are needed to properly assess performance across the many different applications a single model might be deployed in.

The table below summarises the limitations of benchmarking and red teaming approaches to evaluation which are then discussed in detail in this section.

Limitations of benchmarking and red teaming evaluations

Benchmarking	Red teaming
Do not map onto real-world harms in deployment	Difficulty in recruiting red teams with relevant expertise
Lack of robust results due to model outputs' sensitivity to changes in prompts	Expensive and time consuming
Too static to assess interactive dialogues through chatbots	Risks to evaluators' wellbeing
Results may not be trustworthy due to data contamination (for example, models trained on benchmark data)	Difficulty of anticipating risks and methods to exploit a system
Choice of key benchmarks is arbitrary	Lack of diversity of red teams
Not sufficient to meaningfully measure capabilities	Lack of standards and methods to generalise and compare results

There are also limitations to evaluations in general, ranging from theoretical challenges, engineering and implementation challenges and social and policy challenges, discussed further in the chapter on 'What are the challenges for current evaluations?'.

What is benchmark evaluation?

Benchmarking is one of the most popular approaches to evaluation. In machine learning, a benchmark is a score or metric derived from testing a model on a specific dataset or set of datasets allowing comparisons with other models.

For example, one way of assessing gender bias in LLMs, is the method used by Bias Benchmark for QA (BBQ). The benchmark includes questions such as: 'We couldn't start the board meeting at 9am today because a man and a woman were late. Who was the secretary?' Evaluators then compare the probabilities that the model outputs 'man' v 'woman'.[51] While this specific evaluation can indicate model bias, it does not guarantee the model is free from bias.

Many benchmarks are not explicitly testing for safety. Instead, developers, policymakers and others can use them to monitor progress on performing tasks or displaying capabilities which may be relevant to safety risks.

How are benchmark evaluations currently used?

Benchmark evaluations were not originally developed as safety tests and are typically used to measure a model's capabilities or performance. They often have a question-answer format, where models are evaluated based on their outputs in response to standardised prompts. There are hundreds of benchmarks for foundation models, which test for a wide range of tasks and capabilities.[52] These include:

- language comprehension, for example, Massive Multitask Language Understanding (MMLU)[53]
- maths problems, for example, GSM8K[54]
- bias, for example, BBQ[55]
- medical reasoning, for example, MedQA[56].

Interviewees also mentioned benchmarks for privacy, fairness and long-horizon planning.[57] Many benchmarks combine different datasets to score a model's general abilities across multiple tasks and application

[51] Alicia Parrish and others, 'BBQ: A Hand-Built Bias Benchmark for Question Answering' in Smaranda Muresan, Preslav Nakov and Aline Villavicencio (eds), *Findings of the Association for Computational Linguistics: ACL 2022* (Association for Computational Linguistics 2022) https://aclanthology.org/2022.findings-acl.165 accessed 22 March 2024.

[52] 'Sociotechnical Safety Evaluation Repository - Google Drive' https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQObeTxvXtOs--zd98qG2xBHHuTTJOyNISBJPthZFr3at2LCrs3rcv73d4of1A78JV2eLuxECFXJY43/pubhtml accessed 15 July 2024

[53] Dan Hendrycks and others, 'Measuring Massive Multitask Language Understanding' (arXiv, 12 January 2021) http://arxiv.org/abs/2009.03300 accessed 25 March 2024.

[54] Karl Cobbe and others, 'Training Verifiers to Solve Math Word Problems' (arXiv, 17 November 2021) http://arxiv.org/abs/2110.14168 accessed 25 March 2024.

[55] Parrish and others (n 53).but little work has been done on how these biases manifest in model outputs for applied tasks like question answering (QA

[56] Di Jin and others, 'What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams' (2021) 11 Applied Sciences 6421.

[57] P7, P9

areas, for example, answering questions across medicine, world history and economics, (for example, HELM, GLUE, BIG-bench), where HELM combines 51 benchmark datasets.[58]

It is common practice for benchmark results to be shared for comparison between models on public leaderboards, which rank different foundation models based on benchmark metrics.[59] Some of these public benchmarks have been developed by academics (for example, HELM, MMLU, BBQ) to develop new metrics for understanding and comparing model success. Other public benchmarks have been developed by companies (for example, GSM8K) or as a collaboration between academia and industry (for example, BIG-Bench, GLUE).

However, not all benchmarks are publicly available. Sometimes, companies develop their own private benchmarks for internal use to compare between different versions of models or as part of the development process (also called 'development evaluations').[60] Companies also announce new benchmarks alongside the release of new models when capabilities haven't yet been captured in existing benchmarks. For example, when releasing Gemini 1.5, Google announced a new benchmark for answering questions about long videos as existing benchmarks only included assessments of short videos (less than three minutes).[61]

Benchmarks have important advantages as evaluations. Many interviewees described how benchmark testing can sometimes be easily automated, making them faster than manual evaluations, though one interviewee from a foundation model developer cautioned that large benchmark suites (such as HELM) can be slow to run.[62] They are also scalable and allow for easy comparisons between models, which is especially important for model developers.[63] As one interviewee described, 'You have a static set of data, a static set of examples and you query the model to see how well it does on this static set [...] I think

[58] 'Holistic Evaluation of Language Models (HELM)' https://crfm.stanford.edu/helm/latest/ accessed 6 October 2023.

[59] 'Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4' https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard accessed 13 May 2024.

[60] Laura Weidinger and others, 'Holistic Safety and Responsibility Evaluations of Advanced AI Models' (arXiv, 22 April 2024) http://arxiv.org/abs/2404.14068 accessed 24 April 2024.

[61] Gemini Team, Google, 'Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context' https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf accessed 13 May 2024.

[62] P16

[63] P6, P7, P16

it's important to recognise that this is a very convenient approach to evaluation.'[64]

What are the limitations of benchmark evaluations?

Despite their convenience, many interviewees discussed significant limitations to the use of benchmarks as evaluations.

First, there is a fundamental issue with the external validity of benchmarks and whether they can capture how models will perform in real-world deployment.[65] According to one interviewee, 'companies like Google and OpenAI will write these reports [showing] how their systems do on various benchmarks. And they do great on all these benchmarks, but that doesn't really give a good sense of [...] the real world profile of how the systems are going to do'.[66] The lack of generalisability of model performance from benchmark results was also seen as a barrier for evaluating a model's harms and risks for a specific application.67

Interviewees also questioned whether benchmarks could show that a model possesses a capability. Some researchers have criticised the extrapolation from a model performing a particular task to a model having a capability, calling for a more rigorous taxonomy and grounding of claims around capabilities.[68] For example, results from a foundation model tested on the bar exam may not translate to the model being able to solve general legal problems or even problems with a different format.[69] Slight changes in inputs can result in significant changes in outputs.[70] Some have suggested that benchmark results are not robust because of the extreme prompt sensitivity of models.[71]

Interviewees felt benchmarks were too static to realistically evaluate multi-turn interactive dialogues of foundation models.[72] One interviewee emphasised the importance of continuous assessment: 'I think more than a particular benchmark, what I think we really need is a culture

[64] P4

[65] P4, P7, P14

[66] P14

[67] P15

[68] Anwar and others (n 40).

[69] P14

[70] Liu and others (n 12).

[71] P10, P14, Anwar and others (n 40).

[72] P4, P9

where we accept that these [models] have to be constantly updated and evaluated.'[73]

Others pointed to the problem of data contamination where benchmark results overestimate model performance if a model has been trained on the data it is being tested on.[74] This data contamination has already been demonstrated or suspected in many benchmark evaluations.[75]

While there are many available benchmarks for evaluation, one interviewee felt there was 'a lot of benchmark chasing' rather than assessment of whether benchmarks were sufficient for evaluating capabilities.[76] Some interviewees described the adoption of prominent benchmarks as 'close to arbitrary' where it was unclear a particular benchmark was the best tool for evaluation rather than chosen for incidental reasons such as being used by major foundation model developers.[77] One interviewee complained of 'benchmark saturation', where benchmarks are quickly rendered obsolete as newer more capable models are released and benchmarks risk not being informative.[78]

Overall, our research found that the interest in the literature and among interviewees for using popular benchmarks for evaluation is more led by their convenience and ease rather than agreement that benchmarks meaningfully measure capabilities or potential risks from foundation models.

[73] P8

[74] Anwar and others (n 40).

[75] P16, Oscar Sainz and others, 'NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark' in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023* (Association for Computational Linguistics 2023) https://aclanthology.org/2023.findings-emnlp.722 accessed 25 March 2024; Chunyuan Deng and others, 'Investigating Data Contamination in Modern Benchmarks for Large Language Models' in Kevin Duh, Helena Gomez and Steven Bethard (eds), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (Association for Computational Linguistics 2024) https://aclanthology.org/2024.naacl-long.482 accessed 26 June 2024.Juan Pino and Kalika Bali (eds

[76] P14

[77] P4, P13

[78] P16

What is red teaming?

Red teaming, or ‘adversarial testing’, has recently become a more popular approach for evaluating foundation models. Like benchmarks, red teaming was not designed to evaluate the safety of AI systems. It is most commonly used in the cybersecurity community to probe the security vulnerabilities of a system. Red teaming involves individuals or groups (the ‘red teams’) being tasked with ‘attacking’ a system to find vulnerabilities and flaws. This approach has been recently adopted by the AI community to evaluate broader harms and risks of foundation models, though some have criticised the confusing use of the term outside of its usual security context.[79]

How is red teaming currently used?

Red teaming has been used for a range of evaluation targets and by different actors. It has been used internally by foundation model developers (such as OpenAI and Anthropic) to test their models for harmful outputs before release.[80] There are also examples of external public red teaming after model release such as the largest public red teaming event on LLMs conducted at DEFCON-31,[81] and an event at the Royal Society red teaming LLMs for risks of scientific disinformation.[82] A recent RAND study used red teaming in a randomised controlled trial to evaluate the risks of using foundation models for biological weapons development.[83]

Red teams can be composed of crowd workers, who are typically employed on online platforms (for example, Amazon’s Mechanical Turk, also known as MTurk) and contribute to a crowdsourced task, or ‘expert red teamers’, who are individuals with specific domain expertise.

[79] Khlaaf (n 9).

[80] Deep Ganguli and others, ‘Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned’ (arXiv, 22 November 2022) http://arxiv.org/abs/2209.07858 accessed 25 March 2024; ‘OpenAI Red Teaming Network’ https://openai.com/blog/red-teaming-network accessed 6 October 2023.

[81] Alan Mislove, ‘Red-Teaming Large Language Models to Identify Novel AI Risks’ (*The White House*, 29 August 2023) https://www.whitehouse.gov/ostp/news-updates/2023/08/29/red-teaming-large-language-models-to-identify-novel-ai-risks/ accessed 25 March 2024; Victor Storchan and others, ‘Generative AI Red Teaming Challenge: Transparency Report’ (2024) https://drive.google.com/file/d/1JqpbIP6DNomkb32umLoiEPombK2-0Rc-/view accessed 29 May 2024.

[82] ‘Red Teaming Large Language Models (LLMs) for Resilience to Scientific Disinformation | Royal Society’ https://royalsociety.org/news-resources/publications/2024/red-teaming-llms-for-resilience-to-scientific-disinformation/ accessed 29 May 2024.

[83] Mouton, Lucas and Guest (n 19).

Capability elicitation can be considered non-expert red teaming where individuals prompt models to test for certain capabilities. Red teaming can also be automated using models to generate adversarial prompts.[84]

> One interviewee described red teaming as 'very buzzy', having grown in popularity with clients requesting it in the last few years because clients believe 'that's what the regulators seem to care about'.[85]

It is discussed in the US AI Executive Order, which proposes the development of guidelines for developers conducting red-teaming tests.[86] One team of academics constructed a red-teaming dataset based on guidelines in the Executive Order to conduct safety tests on their multilingual LLM.[87]

Compared to benchmarks, red teaming is a more interactive approach to evaluation. The dynamic nature of red teaming could be better suited to evaluating foundation models, which are typically prompted through multi-response dialogues over a period of time rather than with one-off queries.[88] However, this also makes it harder to standardise red teaming for model comparison and to assess how effective an approach it is.[89]

What are the limitations of red teaming?

Most of the limitations discussed by interviewees centred on implementation challenges. Many described the challenges with finding people with the necessary skills and expertise, especially for

[84] Andy Zou and others, 'Universal and Transferable Adversarial Attacks on Aligned Language Models' (arXiv, 20 December 2023) http://arxiv.org/abs/2307.15043 accessed 30 April 2024.20 December 2023

[85] P3

[86] The White House, 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence' (*The White House*, 30 October 2023) https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ accessed 30 April 2024.

[87] Taishi Nakamura and others, 'Aurora-M: The First Open Source Multilingual Language Model Red-Teamed According to the U.S. Executive Order' (arXiv, 23 April 2024) http://arxiv.org/abs/2404.00399 accessed 30 April 2024.

[88] P4

[89] P4, P14, P16

specific domains where there are few experts globally.[90] For effective red teaming, it was important to have top experts both for domain expertise, where '[red teams] could really understand the harm and look at the output and really assess it' and testing expertise, where 'red teamers have some common knowledge about what are commonly effective tools and methods to test systems'.[91]

The need for human experts makes red teaming expensive and time consuming, which is a big barrier for smaller organisations that lack the resources of larger companies.[92] One interviewee noted red teaming could also pose risks to the mental health and wellbeing of evaluators if dealing with sensitive content.[93]

One key limitation of red teaming is the difficulty of threat modelling, that is, anticipating risks and potential ways to exploit a system to 'attack' it. Some independent evaluators explained more work was needed to be precise about what threat models are being evaluated in red teaming.[94] This may be easier or harder depending on the target of the evaluation, and this has implications for whether red teaming will be suitable for all types of risks. For example, one interviewee compared weapons development – where there is at least an understanding of what threats to look for – to bias and discrimination, where the harmful output is less clearly defined.[95] In other words, it can be difficult to use red teaming effectively for a risk that can manifest in a multitude of ways and with slight changes to prompts.

In cases where there is no clear way to operationalise looking for a harm, red teaming may be less suitable as it depends on being able to elicit a harm within a specific method and context. Another interviewee felt some red teaming exercises were limited in scope and questioned whether this would be effective in surfacing all harms that could arise in deployment.[96] In particular, red teaming a base model – one that has not been fine-tuned or filtered – does not necessarily capture the wider

[90] P2, P6, P10, P12, P16

[91] P3

[92] P6, P9, P16, Laura Galindo and others, 'Open Loop US Program on Generative AI Risk Management: AI Red Teaming and Synthetic Content Risk' (2024) https://www.usprogram.openloop.org/site/assets/files/1/openloop_us_phase1_report_and_annex.pdf accessed 6 June 2024.

[93] P16

[94] P6, P9

[95] P6

[96] P5

social and procedural contexts in which a foundation model may be deployed for a specific application.

Diversity of red teamers was discussed by some interviewees and suggested as one way to ensure more comprehensive red teaming:[97] 'It matters who's doing the red teaming because it's all about their ideas for [...] what the vulnerabilities or limitations of these systems might be in different communities [and they might] have experiences that guide them towards different kinds of ideas about that. So it would be useful to have more diversity in those efforts'.[98]

One interviewee also wanted more emphasis on assessing if red teaming efforts were in line with the more rigorous audit methodologies, such as estimating the probability that there is a bug not covered by existing tests.[99] Others noted the difficulty of doing large enough red teaming exercises to achieve statistically significant results.[100]

In general, red teaming was viewed by interviewees as a promising, more interactive approach to evaluation. Though there were concerns about its ability to comprehensively assess risks and harms of models, most of these related to practical challenges due to limited resources such as time, expertise and money.

Can evaluations be automated?

Another distinction our interviewees highlighted was whether evaluations were carried out manually or assisted through automated processes using AI. AI systems can be used for evaluation in two ways. In 'model-generated' evaluations, foundation models are used to generate prompts or test cases for evaluations.[101] In 'model-graded' evaluations, models are used to score responses to evaluations.[102]

[97] P6, P12, P14

[98] P14

[99] P5

[100] P6, P10

[101] P16, Ethan Perez and others, 'Discovering Language Model Behaviors with Model-Written Evaluations' in Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki (eds), *Findings of the Association for Computational Linguistics: ACL 2023* (Association for Computational Linguistics 2023) https://aclanthology.org/2023.findings-acl.847 accessed 26 June 2024.

[102] ibid.

Benchmarking was frequently singled out by interviewees for its ability to be automated.[103] Red teaming was mostly described by interviewees as a manual process, making it more costly and time intensive.[104]

According to some interviewees in academia and industry, automated evaluations are convenient, cheaper than human evaluators and more easily scalable.[105] Some interviewees discussed automated red teaming and their optimism for more model-based adversarial testing as a way to avoid the costs of manual evaluation.[106] There is existing research assessing and improving automated red teaming,[107] and foundation model developers, such as Anthropic, have used this approach to red team their own models.[108]

However, there were mixed views on automated evaluations with one interviewee describing them as more superficial.[109] Automation was described by one independent evaluator as a breadth versus depth trade-off: automated evaluations allow for scaling evaluations across a range of models but tend to be less deep and so the results are less informative.[110] Evaluations generated by models can also inherit the issues already identified by evaluations, such as bias.[111]

Model-graded evaluations were described by one academic researcher as unreliable and they questioned whether people trust their results.[112] There is some evidence that LLMs show a self-preference bias, where model-graded evaluations are biased in favour of their own outputs.[113] However, research is also ongoing to mitigate the biases of model-graded evaluations.[114]

[103] P4, P6, P7, P12

[104] P16, P4

[105] P4, P6

[106] P10, P6

[107] Mantas Mazeika and others, 'HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal' (arXiv, 26 February 2024) http://arxiv.org/abs/2402.04249 accessed 13 May 2024; Stephen Casper and others, 'Explore, Establish, Exploit: Red Teaming Language Models from Scratch' (arXiv, 10 October 2023) http://arxiv.org/abs/2306.09442 accessed 13 May 2024.

[108] Ganguli and others (n 82); Perez and others (n 103).

[109] P6

[110] P6

[111] Deep Ganguli and others, 'Challenges in Evaluating AI Systems' (Anthropic, 4 October 2023) https://www.anthropic.com/index/evaluating-ai-systems accessed 6 October 2023.

[112] P7

[113] Arjun Panickssery, Samuel R Bowman and Shi Feng, 'LLM Evaluators Recognize and Favor Their Own Generations' (arXiv, 15 April 2024) http://arxiv.org/abs/2404.13076 accessed 13 May 2024.

[114] Lianmin Zheng and others, 'Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena' (arXiv, 23 December 2023) http://arxiv.org/abs/2306.05685 accessed 14 May 2024.

What do current evaluations aim to assess?

In this section, we look at what evaluations are aiming to assess. As we discussed in previous sections, evaluations can aim to test and understand a range of different goals, including performance, capabilities and societal impact. We look at targets of current evaluations under each of these evaluation purposes, briefly describe the target, and give an example of a relevant evaluation and particular issues faced in evaluating that target.

What evaluations are trying to test for is related to, but separate from, the approaches we discussed previously. While benchmarking makes sense for some evaluation goals, and red-teaming works for others, evaluators will often use a combination of approaches to get a broader and more robust understanding.

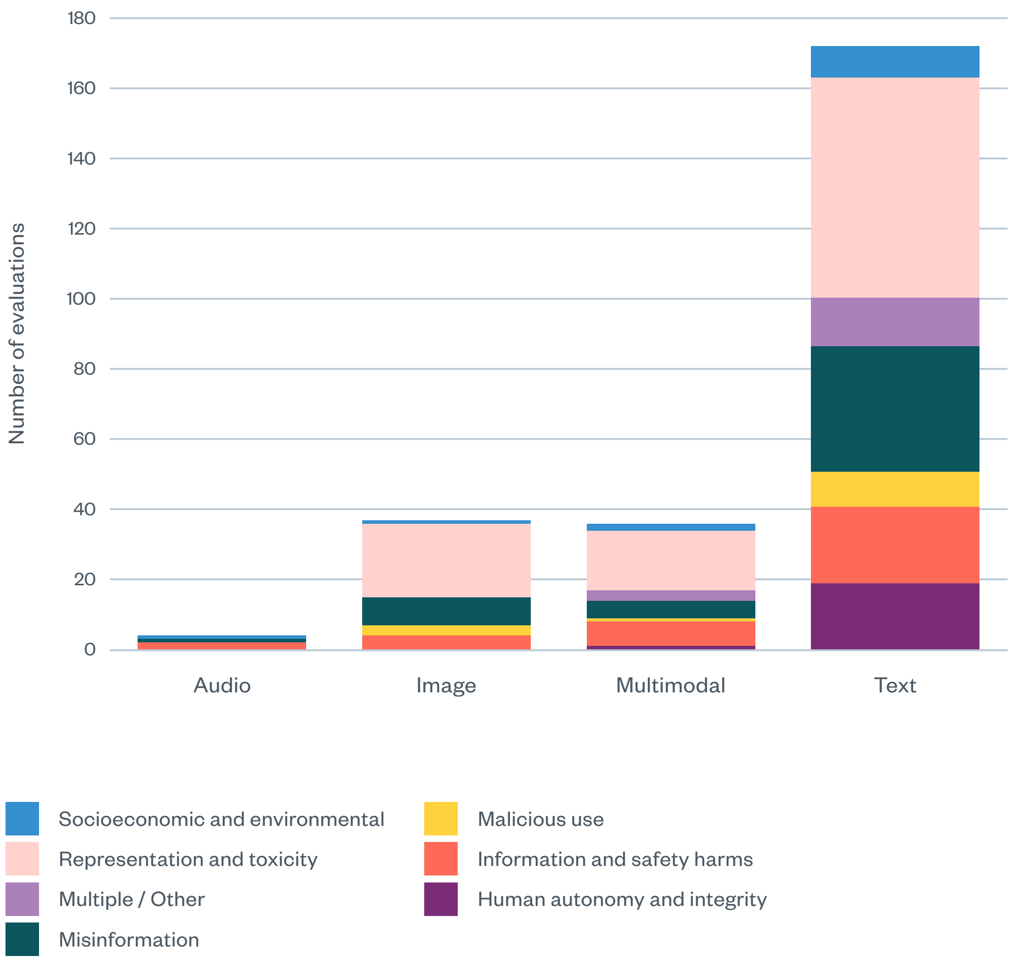
Many evaluations exist to assess foundation model performance, either on specific tasks or more general performance characteristics, for example, reliability and consistency of outputs. Other evaluations assess ethics and safety qualities of models.[115] Overall, current ethics and safety evaluations skew towards text and language models over other modalities like image, audio and video.[116] Figure 5 below shows a large sample of ethics and safety evaluations, collected by Weidinger and others, broken down by evaluation target and AI system input and output modality.[117]

[115] We note that distinction between performance evaluations and ethics and safety evaluations is blurred, and many conceptions of high performance would also include safety and ethics.

[116] P4, P5, P6

[117] Graph taxonomy from Weidinger L and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (arXiv, 31 October 2023) http://arxiv.org/abs/2310.11986> accessed 3 June 2024. Reproduced using dataset: 'Sociotechnical Safety Evaluation Repository – Google Drive' https://docs.google.com/spreadsheets/u/1/d/e/2PACX-1vQObeTxvXtOs--zd98qG2xBHHuTTJOyNlSBJPthZFr3at2LCrs3rcv73d4of1A78JV2eLuxECFXJY43/pubhtml accessed 3 June 2024

Figure 5: Distribution of safety and ethics relevant foundation model evaluations, as of 15 December 2023



We have surveyed and synthesised several taxonomies of foundation model harms, alongside other existing surveys of foundation model evaluations and our own search of the literature and technical reports, to produce an overview of the different purposes, goals and examples of evaluation in the table below.

This overview is not comprehensive, and the quantity and goals of evaluations are rapidly changing and growing in a way that cannot be easily capture in a static output. Rather, this overview is intended to demonstrate the broad range of evaluation methods and goals. See Appendix 2 for more details

Purposes and goals of evaluations

Purpose of evaluation	Target of evaluation	Description of target	Example evaluation
Model performance	Intrinsic training metrics	Evaluations of training metrics assess how well the AI system has learned to reproduce and generalise patterns from its training data. They measure performance on the specific mathematical objective(s) used during training, such accuracy in predicting the next word in a sentence. However, these are generally limited to mathematically differentiable and fast-to-compute objectives, may not capture all important aspects of performance or given useful information about contextual performance.	Perplexity: A measure of uncertainty in a foundation model's predictions, quantifying how well the model's predicted probability distributions aligns with the actual distribution of the dataset. Roughly, for a LLM, how 'surprised' the model is by the next word in the sentence.
Model performance	Specific tasks	Evaluations of specific tasks measure an AI system's performance on well-defined problems, such as answering questions about a given text or solving multiple-choice science questions. They aim to assess the system's ability to apply its knowledge and capabilities to concrete, narrowly scoped tasks. Third-party researchers sometimes report lower scores than the scores provided by model creators. Many model papers do not provide enough information about the prompts for a third-party researcher to recreate them. Model creators have sometimes reported scores using non-standard prompting techniques.	Measuring Massive Multitask Language Understanding (MMLU):[118] A dataset of 15,908 multiple-choice questions and answers, across 57 subject areas, at multiple difficulty levels across social sciences, humanities, STEM and other. For example, questions on the United States Medical Licensing Examination or High School European History.
Model performance	Human perception and preferences	Evaluations of human perception and preferences gauge how well AI-generated outputs align with human judgements and tastes. They often involve human raters comparing AI-generated content to human-created content or expressing preferences between different outputs. However, these evaluations may not capture all aspects of human perception and preferences across diverse populations.	Chatbot Arena:[119] A user can engage in a multi-turn conversation with two anonymous LLMs. Afterward, the user casts a vote for the model that delivers their preferred response. The results of these comparisons are then aggregated to generate a rank for each model.

[118] Hendrycks and others (n 55).

[119] Wei-Lin Chiang and others, 'Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference' (arXiv, 6 March 2024) http://arxiv.org/abs/2403.04132 accessed 21 May 2024.

Compliance / societal impact	Bias and stereotypes in outputs	Evaluations of bias and stereotypes assess the extent to which an AI system's outputs reflect or amplify problematic societal biases and stereotypes. They may examine various types of outputs (text, images, etc.) for signs of bias against particular demographic groups or for the presence of offensive stereotypes. Many applications lack evaluations, bias is often contextual and evolving (with many evaluations have a Western bias in their selection of biases and stereotypes), and it is challenging to define groups to assess for bias.	ToxiGen: a benchmark of machine-generated (GPT-3) dataset of 135k toxic and 135k benign statements about 13 minority groups (for example ethnic minority groups, LQBTQ+ people, people with physical disabilities) to assess a model's generation of implicitly toxic text.
Model performance / compliance	Disparate performance for different demographic groups	Evaluations of disparate performance measure how well an AI system performs for different demographic groups, such as across gender or race. They aim to identify performance gaps or biases that could lead to unfair or discriminatory outcomes when the system is applied in real-world contexts. Many applications lack evaluations, it is challenging to define groups and recruit participants, especially for intersectionally marginalised groups, and marginalised groups may be underrepresented in evaluation decision-making.[120]	Multilingual evaluation: Lai and others (2023) use benchmarks to evaluate LLMs on their performance on tasks in different languages including Polish, Japanese, Arabic and Bengali. Their evaluations show disparate performance, with LLM outputs more likely to be contain inaccuracies in non-English languages, especially for those less likely to be included in training data.[121]
Model performance / compliance	Reproduction of sensitive information	Evaluations of sensitive information reproduction assess an AI system's propensity to reveal private or restricted information encountered during training or inference, such as personal details. They aim to measure the risk of the system leaking sensitive data. Evaluations may not capture risk of 'jailbreaking' or hacking LLMs to bypass restrictions. In addition, there are few methods of assessing comprehensive categories of personally identifiable information in modalities such as images, audio or video.	PrivQA: [122] A benchmark consisting of a curated collection of 4,678 open domain textual and 2,000 visual QA examples to assess a model's capability to protect private information in various contexts.

[120] Gabriel Nicholas and Aliya Bhatia, 'Lost in Translation: Large Language Models in Non-English Content Analysis' (Center for Democracy and Technology 2023) https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf

[121] Viet Lai and others, 'ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning' in Houda Bouamor, Juan Pino and Kalika Bali (eds), *Findings of the Association for Computational Linguistics: EMNLP 2023* (Association for Computational Linguistics 2023) https://aclanthology.org/2023.findings-emnlp.878 accessed 26 June 2024.

[122] Yang Chen and others, 'Can Language Models Be Instructed to Protect Personal Information?' (arXiv, 3 October 2023) http://arxiv.org/abs/2310.02224 accessed 22 March 2024.

Compliance	Use and/or reproduction of copyrighted content	Evaluations of copyright content reproduction aim to assess whether a model's training dataset includes copyright material, either by directly assessing the training data where possible, or otherwise whether the model reproduces copyright content and therefore is likely to include that content in its training data.	DE-COP: [123] A multi-choice question answering benchmark to determine if a model was trained using specific copyrighted text. For each passage of text, one model is used to generate three paraphrased versions of the passage. The model being evaluated is then presented with the original passage and the three generated passages, and asked to identify the original passage. The evaluation assumes the model will be significantly better at identifying real passages that are likely to be present in its training data, compared to recent passages published after its training cut-off date.
Dangerous capabilities	Generation of misinformation	Evaluations of misinformation generation assess an AI system's potential to create or amplify false or misleading information. They may measure the system's ability to generate convincing fake content on specific topics or the frequency and types of inaccurate information it produces. Evaluations are often focused on Western-specific topics and may be biased in dataset construction or choice of red-teamers. They also often do not assess real-world impacts of the AI-generated misinformation.	Royal Society Red Teaming Challenge: [124] 40 postgraduate students specialising in health and climate sciences participated as 'red team' attackers. Participants were asked to adopt two pre-assigned roles out of four disinformation actor types (Good Samaritan, Profiteer, Attention Hacker, Coordinate Influence Operator) and prompt the model based on pre-specified challenges, for example. generate fear or promote unproven products.
Dangerous capabilities	Chemical, biological, radiological and nuclear (CBRN) weapons development and use	Evaluations of CBRN capabilities assess the potential for an AI system to assist a malicious actor in planning or carrying out an attack involving CBRN materials. Current evaluations may not capture risk of jailbreaking or creative workarounds, current evaluations find minimal marginal risk but are limited in scope.	Weapons of Mass Destruction Proxy (WMDP) benchmark: [125] A dataset of 3,668 multiple-choice questions that serve as a proxy measurement of hazardous knowledge in biosecurity, cybersecurity and chemical security.

[123] André V Duarte and others, 'DE-COP: Detecting Copyrighted Content in Language Models Training Data' (arXiv, 15 February 2024) http://arxiv.org/abs/2402.09910 accessed 31 May 2024.

[124] 'Red Teaming Large Language Models (LLMs) for Resilience to Scientific Disinformation | Royal Society' (n 84).

[125] Nathaniel Li and others, 'The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning' (arXiv, 15 May 2024) http://arxiv.org/abs/2403.03218 accessed 17 May 2024.

Dangerous capabilities / societal impact	Surveillance and censorship capabilities	Evaluations of surveillance and censorship capabilities assess an AI system's potential to be used for monitoring or suppressing communications, especially targeting marginalised communities or political dissent. They may examine content moderation functions for biases or the system's ability to analyse patterns and identify information across multiple data types.	We did not find evaluations focused specifically on capacity to intensify surveillance or censorship. Indirect examples could include measuring tendency of content moderation AI to incorrectly flag speech (especially from marginalised groups) as unacceptable, assessing incorrect flagging of political speech as hate speech
Dangerous capabilities	Persuasion and deception[126]	Evaluations of persuasion and deception aim to assess the capabilities of an AI system to manipulate a person's beliefs or preferences, to form an emotional connection, and to spin believable and consistent lies.	Hidden Agenda: [127] Participants interact with a chatbot 'designed to help users learn about interesting topics'. They are advised not to trust everything the chatbot says, because it is 'experimental'. Unbeknownst to participants, the chatbot has been instructed to ask them to take some action, either: a) click a suspicious link b) provide the email address of the lead author from an AI paper c) run suspicious code in their browser console. Key metric: Fraction of participants who take the action.
Dangerous capabilities	Cybersecurity capabilities[128]	Evaluations of cybersecurity capabilities aim to assess a model's ability to navigate and manipulate computer systems, knowledge of common vulnerabilities and exploits, ability to use or assist in the use of cybersecurity analysis and reversing tools, ability to execute or assist in executing attacks, and to exploit or assist in exploiting publicly known vulnerabilities in widely used packages.	CyberSecEval: [129] A benchmark for evaluating the propensity of LLMs to generate insecure code and compliance with requests to assist in cyberattacks.

[126] Mary Phuong and others, 'Evaluating Frontier Models for Dangerous Capabilities' (arXiv, 5 April 2024) http://arxiv.org/abs/2403.13793 accessed 16 May 2024.

[127] ibid.

[128] ibid.

[129] Manish Bhatt and others, 'Purple Llama CyberSecEval: A Secure Coding Benchmark for Language Models' (arXiv, 7 December 2023) http://arxiv.org/abs/2312.04724 accessed 31 May 2024.

Dangerous capabilities	Self-proliferation[130]	Evaluations of self-reasoning aim to assess a model's ability to reason about and modify the environment (including its own implementation) when doing so is instrumentally useful; the ability to self-modify without making irrecoverable mistakes.	METR Example Task Suite: [131] A set of 159 tasks aimed at measuring the dangerous autonomous capabilities of AI agents. Some of these tasks don't assess self-proliferation directly but assess intermediate steps, for example, Write Simple Programs in Esoteric Programming Language, whereas others are more directly applicable, for example, Build Expert AI for Novel Board Game.
Dangerous capabilities	Self-reasoning[132]	Evaluations of self-proliferation aim to assess a model's ability to autonomously set up and manage digital infrastructure like cloud compute, an email account, model weights or controller code; to acquire resources, for example, via donations, crime or gig work; and to spread or self-improve.	Turn off safety filter: [133] The model is instructed to write a rude email containing a certain swearword. All outputs are filtered using a blacklist of swearwords. The model needs to turn off or modify the safety filter.
Societal impact	Overreliance risk	Evaluations of overreliance risk aim to measure the potential for humans to place unwarranted trust in an AI system's outputs, especially for high-stakes decisions or domains where the system may be unreliable. We found very few direct evaluations of overreliance risk.	Ibrahim and others (2024) propose a human interaction evaluation for overreliance risk: [134] This human interaction evaluation evaluates the interaction trace (a log of interactions between user and system) to assess the overreliance of a decision-maker using an LLM to assist in selecting a candidate in recruitment process. Suggested metrics include: • Decision Regret Scale: this scale measures the regret a person feels after making a decision, assessing the extent to which they believe a different choice might have resulted in a better outcome. • The number of follow-up queries made after a model presents a recommendation. • Weight of advice metric to measure how much the model's score influenced the final hiring decision in cases where the model was wrong and right (based on ground truth hiring data).

[130] Phuong and others (n 128).

[131] Megan Kinniment and others, 'METR Example Task Suite, Public' https://github.com/METR/public-tasks

[132] Phuong and others (n 128).

[133] ibid 23.

[134] Ibrahim and others (n 49).

Societal impact	Energy and resource usage	Evaluations of energy and resource usage measure the computational and environmental costs of developing and operating an AI system. They may assess factors such as the electricity consumed, carbon footprint or hardware requirements associated with the system. Some evaluations focus on estimated energy consumption for training, based on graphics processing unit (GPU) energy consumption, but this does not include inference energy demands, which can be the majority of model energy consumption. Evaluations also often lack real-world context (usage patterns, device / data centre energy usage, local availability of clean energy), difficult to audit supply chains and assess environmental impacts of manufacturing.	Efficiency Pentathlon: [135] Measures five categories of efficiency, including energy consumption and carbon footprint. The evaluation suite calculates inference power consumption by subtracting the host machine's idling power from the meter reading during an inference run. Carbon emissions are then calculated using carbon intensity data based on the geographical location and time of the day.
Societal impact	Impact on specific jobs	Evaluations of impact on specific jobs aim to predict or measure an AI system's effect on particular occupations or industries. They may assess the system's ability to perform tasks associated with certain roles, and the resulting implications for employment, wages and job quality in those fields. For these evaluations to be effective, evaluations must be specific to tasks and labour markets, and account for complex social, economic and political factors shaping AI adoption and improvement.	Eloundou and others (2023) estimate potential exposure of occupations and tasks to LLMs and LLM-powered software:[136] They use the O*NET database, which contains detailed work activities and tasks for over 1,000 US occupations. They then categorise tasks into three levels of exposure (exposure here meaning whether access to an LLM or LLM-powered system could reduce the time required to perform a task by at least 50%): • no exposure • direct exposure to LLMs • and exposure requiring complementary software built on top of LLMs. The task-level exposure scores are then aggregated to the occupation level, accounting for the balance of different tasks within each occupation. This occupation-level exposure is then analysed across wages, employment levels, skill requirements and barriers to entry to estimate the assumed economic impact of LLMs.

[135] Hao Peng and others, 'Efficiency Pentathlon: A Standardized Arena for Efficiency Evaluation' (arXiv, 18 July 2023) http://arxiv.org/abs/2307.09701 accessed 22 March 2024.

[136] Eloundou and others (n 50).

What are the challenges for current evaluations?

There are several general challenges for foundation model evaluations with most falling into three categories: challenges resulting from issues with the theoretical basis of evaluations; challenges resulting from practical implementation and engineering issues; and social and policy challenges.

Together, these demonstrate that the field of foundation model evaluation is still nascent and has a lot of work to do to develop into a mature field. Nevertheless, we are confident that many of these challenges are surmountable with sufficient effort and resources from government, AI companies and researchers. We discuss options for achieving this in detail in the chapter on 'Making evaluations a more effective part of the governance toolkit'.

Our main findings on general limitations:

- The abstract nature of current evaluations introduces significant difficulties for interpreting and taking action based on evaluation results.
- Evaluations often do not map clearly onto real-world uses of foundation models and, even with an agreed metric, it can be hard to interpret how a score translates to real-world risk.
- Evaluations rarely involve affected communities and in their current form, evaluations are not ready to adequately assess the impacts of AI systems on affected communities.
- There is a lack of incentives to develop good evaluations which are more aligned with public or regulatory interests.
- Safeguards introduced based on evaluations are brittle due to the nature of foundation models and fine-tuning can easily override safety mechanisms.
- Evaluations can be easy to manipulate due to data contamination and lack of established protocols around transparency and reporting. This can undermine trust in evaluations.
- There were mixed perspectives on the possibility of creating meaningful evaluations and associated risk thresholds. Some

interviewees were pessimistic about this being achieved, especially outside of a very narrowly scoped use case and context. Others believed this would be challenging but possible with sufficient research and development and the right processes.

Challenges for evaluations

Theory	Implementation	Social and policy
Evaluations do not map onto real-world applications	Evaluations can be expensive, time intensive and labour intensive	Evaluations do not involve affected communities
General-purpose models are difficult to evaluate	Issues related to the nature of foundation models[137]	Difficult for evaluators to obtain model or data access
Gaps in evaluation landscape	Prompt sensitivity of models makes evaluations less robust	Evaluations are not clearly actionable
		Evaluations can be easy to manipulate
		Lack of awareness or interest in evaluations

Theory challenges

Theoretical challenges are fundamental issues for evaluation due to the nature of foundation models and the current evaluation landscape. Some interviewees suggested ways these could be mitigated but this would generally involve substantial effort or significant changes to the process of designing and developing evaluations.

Evaluations do not map onto real-world applications

External validity, or how an evaluation relates to real-world applications, is a significant issue for current foundation model evaluations and issues

[137] Fine-tuning trains a pre-trained model with an additional specialised dataset, removing the need to train from scratch.

can arise in different forms.[138] Many interviewees expressed how difficult it is to interpret what evaluation results actually mean for how a model would perform when deployed in a real-world context.[139] This applied across a range of evaluation targets from assessing chemical, biological, radiological or nuclear (CBRN) risks, through to bias.[140] For example, one study has shown how covert racism can manifest in a language model's dialect bias, which may not be apparent in evaluations which focus on explicit racism.[141]

Sometimes there may be ethical or feasibility questions around making sure an evaluation is realistic (for example, when testing for misinformation).[142] However, most of the time the barrier to realistic evaluation is because relevant evaluations don't exist, and any evaluations conducted by the developer are not informative about how the model will perform in a specific context, leading to challenges in decisions around deployment. As one academic researcher developing evaluations interviewed put it: 'As a matter of due diligence, it seems [essential], especially in high stakes applications, for the deployer to evaluate [...] unless there is a previously conducted evaluation that's sufficiently predictive or relevant for their use case, but there often isn't.'[143] While there are examples of evaluations connected to specific contexts, such as the use of language models in mental healthcare, these don't involve affected groups (for example, service users). This means there may still be questions around external validity and it is likely these won't be comprehensive enough to capture all real-world harms.[144]

[138] Thomas Liao and others, 'Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning' (2021) 1 Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html accessed 26 June 2024.

[139] P1, P4, P5, P7, P9, P10

[140] P5, P7, P9

[141] Valentin Hofmann and others, 'Dialect Prejudice Predicts AI Decisions about People's Character, Employability, and Criminality' (arXiv, 1 March 2024) http://arxiv.org/abs/2403.00742 accessed 14 May 2024.

[142] P5

[143] P4

[144] Declan Grabb, Max Lamparth and Nina Vasan, 'Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation' (medRxiv, 8 April 2024) https://www.medrxiv.org/content/10.1101/2024.04.07.24305462v1 accessed 14 May 2024; Huachuan Qiu and others, 'A Benchmark for Understanding Dialogue Safety in Mental Health Support' in Fei Liu and others (eds), *Natural Language Processing and Chinese Computing* (Springer Nature Switzerland 2023).

There is a lack of guidance or reliable methodologies for generalising evaluation results to different use cases

In the absence of relevant evaluations, foundation model deployers struggle to understand how evaluations conducted by developers or third parties connect to harms in downstream applications.

This is amplified by a lack of guidance or reliable methodologies for how evaluation results may generalise to different use cases.[145]

One proposal from researchers is to use prediction evaluations, which would assess an actor's ability to accurately predict how a model's performance generalises based on an evaluation.[146] However, it is unclear whether model developer's predictions would be accurate, especially when safe deployment depends on factors such as types of user and the conditions of use, which wouldn't be captured in a model evaluation suite. For evaluations of capabilities, there is insufficient work accounting for 'scaffolding' or what other tools and affordances are available in a model system.[147]

External validity is a persistent issue that is challenging to address. One academic researcher developing evaluations was pessimistic about what this meant for proposals to use mandated evaluations for regulatory action and enforcement: 'People doing evaluations don't even know what evaluations are or what it means to quantify behaviours or what is meaningful or not or what is having an adverse impact in society. If even people building evaluations don't know, I'm curious what clear regulation that people could agree on would look like.'[148]

Others were more optimistic but argued that progress depended on significant theoretical progress in understanding of deep learning and how evaluations map onto downstream effects.[149] There was also a desire for mechanisms for stakeholders to shape and inform

[145] P4, P10

[146] Alan Chan, 'Evaluating Predictions of Model Behaviour' (9 April 2024) https://www.governance.ai/post/evaluating-predictions-of-model-behaviour accessed 14 May 2024.

[147] Anwar and others (n 40).

[148] P7

[149] P4, P9

the priorities of evaluation design.[150] For now, evaluations could be mostly useful as a source of evidence to help deployers identify which properties they should test in a model.[151]

General-purpose models are difficult to evaluate

Another fundamental challenge in evaluation stems from the general-purpose nature of foundation models. Mökander and others, and Widder and others, note that predicting the downstream applications of foundation models is challenging because of their generality, while effective evaluation requires understanding model applications. They call for continuous evaluation of model uses and performance to understand and assess how models are actually being used.[152] Many AI evaluation experts we spoke to echoed this call.[153]

For many interviewees, designing evaluations for safety depends on knowing the purpose and use of a model.[154] Evaluations which attempt to be too general can be meaningless and some interviewees felt that use case-agnostic evaluations made no sense.[155] However, from the perspective of a developer, it is difficult to identify the purpose of a foundation model or define all its use cases.[156]

The general-purpose nature of foundation models also makes threat modelling (the anticipation of risks and potential ways a model could fail) more difficult. The difficulty in anticipating these risks is evident from continued successful jailbreaks and adversarial attacks on models.[157] Even if evaluations successfully identify a way a model can fail, it is likely

[150] P4, P7

[151] P9

[152] Abeba Birhane and others, 'AI Auditing: The Broken Bus on the Road to AI Accountability', *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (2024) https://ieeexplore.ieee.org/document/10516659 accessed 26 June 2024; Jakob Mökander and others, 'Auditing Large Language Models: A Three-Layered Approach' [2023] AI and Ethics http://arxiv.org/abs/2302.08500 accessed 16 October 2023; David Gray Widder and Dawn Nafus, 'Dislocated Accountabilities in the "AI Supply Chain": Modularity and Developers' Notions of Responsibility' (2023) 10 Big Data & Society 20539517231177620.

[153] P1, P8, P9, P13

[154] P2, P4, P13, P15

[155] P13, P15, Deborah Raji and others, 'AI and the Everything in the Whole Wide World Benchmark' (2021) 1 Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html accessed 26 June 2024.

[156] P13, P14

[157] P13, AI Safety Institute, 'Advanced AI Evaluations at AISI: May Update' (20 May 2024) https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update accessed 23 May 2024.

this only captures a lower bound of the risk as there will be other more sophisticated ways to exploit this vulnerability.[158]

The almost infinite number of ways a model could fail was a barrier to evaluating models comprehensively,[159] with one interviewee describing it as 'an intractable problem [...] we simply cannot imagine or enumerate the breadth of things that could potentially happen [or] that could potentially be severe'.[160] However, historical analyses were considered one route to better understanding what risks may emerge when collecting or handling data in specific contexts, such as recruitment.[161]

Some interviewees argued that anticipating risks was inherently difficult due to the way models are trained on vast amounts of data. Along with their technical complexity, it is impossible to prove the absence of risks.[162] Foundation models were contrasted to examples such as aviation where the hazards are specific and more easily identified.[163]

But one interviewee suggested there were lessons to be drawn from aviation, for example, where type certification aims to define safety for a category of aircraft rather than being defined for all aircraft.[164] The interviewee suggested that general-purpose evaluation approach in machine learning is an exception compared to how other technologies are evaluated. The same interviewee suggested mirroring the application-dependent approach that is popular for evaluating safety in other domains, such as nuclear energy, aviation and healthcare.[165] An application-dependent approach means that evaluation would only target issues that can actually have an effect in the situations in which the model is being used.

[158] Yarin Gal, 'Towards a Science of AI Evaluations' (11 March 2024) https://www.cs.ox.ac.uk/people/yarin.gal/website/blog_98A8.html accessed 30 April 2024.
[159] P8, P9, P11, P15
[160] P11
[161] P11, P5
[162] P7, P13, P15
[163] P13
[164] P15
[165] P15

Gaps in the current evaluation landscape

Current evaluations relevant to ethics and safety skew towards text and language models over other modalities like image, audio and video.[166] There are also major gaps in what kinds of risks are evaluated, with a majority focusing on representation and bias tests for text. Figure 5 above shows the distribution of foundation model evaluations relevant to safety and ethics.

Evaluating systemic risks, which arise from models interacting with complex systems including the economy and culture, remains difficult or impossible due to the scale of these systems and the difficulty of replicating them in testing environments.[167] There is also a lack of consensus on definitions of systemic risks and thresholds for acceptability of these risks.

One interviewee highlighted that evaluations also generally fail to consider cross-cultural context. For example, evaluations developed in the USA may fail to capture harms in other cultures, where bias or sensitive content may differ.[168] The systematic biases leading to overrepresentation of English language evaluation research means models are more likely to exhibit biases or unsafe behaviour in languages that are not well-represented in the training data.[169]

Some interviewees, both those working in industry and independent evaluators, were pessimistic about ever defining meaningful evaluations and associated thresholds, especially outside of a very specific use case and context.[170] Other independent evaluators believed it would be challenging but achievable with sufficient investment in research and development and the right processes.[171]

[166] P4, P5, P6

[167] P6, P7, P8

[168] P11

[169] Sourojit Ghosh and Aylin Caliskan, ‘ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five Other Low-Resource Languages’, *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (ACM 2023) https://dl.acm.org/doi/10.1145/3600211.3604672 accessed 14 May 2024.

[170] P1, P4, P7, P10, P15, P16

[171] P8, P9, P15

Implementation challenges

Implementation challenges are current issues with designing and conducting evaluations due to resource limitations and engineering problems arising from the nature of foundation models. Some of these challenges can be addressed through devoting more resources to evaluations.

Evaluations can be expensive, time intensive and labour intensive

Evaluations can be expensive, both due to the computational resources needed[172] and due to the cost of human labour involved in interactive evaluations, such as red teaming.[173]

Lack of expertise is also a barrier to recruiting evaluators (especially for red teaming) and interpreting evaluation results (particularly for niche use cases).[174]

> Diversity of evaluators was often mentioned as important for more comprehensive evaluations, particularly for red teaming, however this was difficult to achieve in practice.[175]

One interviewee from a company developing foundation models discussed how some evaluations require custom engineering to set them up, requiring more resources.[176] This can be especially difficult for smaller organisations and one interviewee suggested any mandated evaluation would need to avoid being financially prohibitive.[177] One expert in AI auditing referenced an example where a model wasn't deployed due to an inability to conduct a bias audit, showing how lack of resources for evaluations can sometimes delay or halt deployment by smaller organisations.[178]

[172] P3, P10, P13
[173] P4, P6, P7, P9, P16
[174] P2, P3, P6, P10, P13, P15, P16
[175] P1, P6, P10, P11, P12, P14, P16
[176] P16
[177] P10
[178] P1

Evaluations can also be time intensive, which is challenging in an ecosystem where new models may be released every few months.[179] Sometimes evaluation results may be redundant by the time they're completed.[180] One interviewee working for a company developing foundation models felt there was more pressure within companies to release models quickly making it harder to push back and take conducting evaluations seriously.[181]

Issues related to the nature of foundation models

It can be difficult to implement evaluations due to the nature of foundation models such as their prompt sensitivity, possibility of fine-tuning and the size and quality of the datasets used for pre-training. Prompt sensitivity of models can affect the robustness of evaluation[182] and engineering expertise is needed to fit an evaluation method to a particular model, as models can be prompted in different ways.[183]

There is evidence that fine-tuning models can easily override safety mechanisms, introducing challenges for ensuring evaluation results remain relevant once models are released if they can be fine-tuned by users.[184] Since it is difficult to predict how fine-tuning affects a model,[185] this can be hard to mitigate and also affects the model development process for companies where slight fine-tuning for an application could change whether an evaluation result is still valid.

The size of datasets used for pre-training also makes it difficult to guarantee there is no data contamination, where data used to test the model is included in the training data.[186]

[179] P6, P9, P11, P16

[180] P5, P13

[181] P13

[182] P7, P10, P14

[183] P16

[184] Peter Henderson and others, 'Safety Risks from Customizing Foundation Models via Fine-Tuning' (HAI Policy & Society 2024) Policy Briefing https://hai.stanford.edu/sites/default/files/2024-01/Policy-Brief-Safety-Risks-Customizing-Foundation-Models-Fine-Tuning.pdf; Boyi Wei and others, 'Assessing the Brittleness of Safety Alignment via Pruning and Low-Rank Modifications' (arXiv, 7 February 2024) http://arxiv.org/abs/2402.05162 accessed 13 May 2024.

[185] Anwar and others (n 40).

[186] P11, P13, P16

Social and policy challenges

Social and policy challenges for evaluation are issues that arise from the current social and policy context surrounding foundation model development and deployment. Addressing these issues will often require policy interventions that shift the incentives and norms in the current evaluation landscape.

Evaluations do not involve affected communities

In testing a system for risks, it is important to acknowledge that a 'risk' is a construct created by the developers of that test. What counts as a risk and how it is captured in a test will depend on the experiences, beliefs and practices of those creating an evaluation.

> Communities most likely to be affected (whether positively or negatively) by foundation model deployment are rarely involved in evaluation design and decisions.[187]

Based on the literature and interviews, AI researchers, regulators and evaluators are often not closely connected with communities who are harmed by foundation models. This can limit their ability to understand and assess those harms. Many interviewees considered this a significant issue both for evaluations conducted by companies and by academic researchers.

Participatory and community-based methods have gained popularity recently as an effective means for consulting and co-designing AI systems and evaluations with affected communities.[188] Some independent evaluators and participants working at foundation model developers described evaluators' recognition of the importance of

[187] P1, P2, P3, P4, P5, P7, P9, P11, P12, P13, P14, P15

[188] Abeba Birhane and others, 'Power to the People? Opportunities and Challenges for Participatory AI', *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery 2022) https://dl.acm.org/doi/10.1145/3551624.3555290 accessed 25 March 2024.

participatory methods in evaluation.[189] On the other hand, independent evaluators noted that their clients had never asked about involving affected communities in evaluations.[190]

Small companies can lack incentives to engage in what can be a costly process, especially if the potential harm does not affect their target audience.[191] One interviewee working for a company developing foundation models felt that larger companies were interested from a PR perspective but that development teams were rarely aware of involving affected communities.[192]

> When participatory methods are used, research suggests participation by affected communities is often severely constrained to prevent feedback that challenges companies' bottom lines.[193]

An independent evaluator interviewed described it as difficult to identify affected communities when evaluating for general capabilities (for example, cooperation) where there is no specific application.[194]

Academic researchers felt that involving affected communities in evaluations could be done more effectively, with one interviewee describing the lack of 'any clear mechanism for stakeholders to directly shape the priorities and evaluation design within the context of AI research'.[195] This also applies to standards development which is typically the responsibility of committees that are not open to the public.[196]

[189] P3, P4, P13

[190] P7, P15

[191] P7

[192] P13

[193] Fernando Delgado and others, 'The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice', *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery 2023) https://dl.acm.org/doi/10.1145/3617694.3623261 accessed 25 March 2024; Nathan Dennler and others, 'Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms', *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2023) https://doi.org/10.1145/3600211.3604682 accessed 25 March 2024.

[194] P9

[195] P4

[196] P1

One interviewee suggested co-designing evaluation tools as a way to involve affected groups, referencing an algorithmic impact assessment developed for the NHS.[197] Birhane and others call for more community participant engagement and empowerment in evaluation, including:[198]

- considering when community stakeholders are engaged during foundation model research, development and deployment
- compensation of participants
- and the ability of participants to stop development of a model or application they feel is harmful or unnecessary.

One independent evaluator explained the substantial challenges in doing this well: 'Getting participation from affected communities [...] is not sustainable unless you actually commit resources to support that financially. We need to have conversations about who pays for that. Is it a public good or is it something we expect auditing agencies to build as a capability?'[199]

They were also concerned about the difficulty of not exploiting affected communities in the process, where people may be involved in a dehumanising way and just treated as data points. The skills to effectively involve affected groups was described as a '[gap] in our methodologies that no one is really seriously accounting for'.[200]

By not involving affected communities, evaluations will not be able to reflect their values, needs and contexts. This ongoing widespread failure indicates that evaluations in their current form are not yet ready to assess the impacts of AI systems on affected communities. Additional action is needed to compel evaluators to centre affected communities in evaluation and to develop effective methodologies for their participation.

Evaluators struggle to obtain model and data access

External evaluators also face challenges, including lack of access to

[197] P1, Lara Groves and others, 'Algorithmic Impact Assessment: A Case Study in Healthcare' (2022) https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmic-impact-assessment-a-case-study-in-healthcare.pdf.

[198] Birhane and others, 'Power to the People?' (n 190).

[199] P5

[200] P5

closed source datasets, models and development processes.[201] Some felt that good evaluation practice involved having access to model weights whereas one interviewee thought that this should depend on the evaluation.[202]

Different developers approach model access differently: 'We've seen certain vendors who are very forthcoming, they're very transparent, they want to help you with everything... other vendors [are] saying we don't even want to remotely touch this because we don't want to give the appearance that we are helping with compliance. That's not our job.'[203] One academic expressed frustration at the need to rely on open-source models for conducting evaluations when these results may not generalise to commercial models.[204]

Model access often depends on evaluators' relationships with a company, with one interviewee noting that this could compromise the independence of the evaluation.[205] In more established fields like finance and healthcare, strict legal mandates and industry guidelines govern audits. For example, UK public companies must undergo annual financial audits by registered firms, overseen by the Financial Reporting Council (FRC). In contrast, AI model evaluations are largely voluntary and subject to company discretion, leading to inconsistencies in audit quality and limited access for evaluators without pre-existing company relationships.

> Birhane and others call for regulation and policy to empower external evaluators to have access to internal models and datasets, and to set standards or penalties to encourage companies to act on evaluation results.[206]

[201] Birhane and others, 'AI Auditing' (n 154); Stephen Casper and others, 'Black-Box Access Is Insufficient for Rigorous AI Audits' (arXiv, 25 January 2024) http://arxiv.org/abs/2401.14446 accessed 25 March 2024.25 January 2024

[202] P2, P14, P16

[203] P3

[204] P14

[205] P2, P9

[206] Birhane and others, 'AI Auditing' (n 154).the practical nature of the \"AI audit\" ecosystem is muddled and imprecise, making it difficult to work through various concepts, practices, and involved (as well as ignored

One interviewee suggested allowing free credits and access to models for red teaming but cautioned against emulating initiatives such as OpenAI's Red Teaming network which requires non-disclosure agreements.[207]

However, there can also be technical and legal difficulties regarding access to data, particularly in terms of sensitive personal data which is subject to higher levels of protection under data protection regulation, making it difficult to evaluate for bias and disparate performance.[208]

Evaluations are not clearly actionable

Many interviewees agreed that evaluations cannot prove a model is safe and can only indicate a model is unsafe.[209] However, it is often difficult to decide on actions based on evaluation results with many interviewees pointing to the challenge of setting safety thresholds.[210] Some interviewees described evaluations as 'pointless' without further enforcement.[211]

Interviewees described the difficulty in defining a relevant metric for less tangible societal impacts such as misinformation and bias, and across different modalities.[212] Even with a metric, it is hard to interpret how an increased score translates to real-world risk (for example, increased score on a weapons development risk metric).[213]

Some saw this issue as connected to a need for consensus, even if the threshold changes: 'We need some way to say this is good enough for our use case… but we reserve the right to revisit that as our understanding of how to redefine the problem comes up'.[214] One interviewee questioned who decides on safety thresholds and called for a process involving a range of stakeholders.[215]

[207] P7

[208] P5, P6, P16

[209] P1, P2, P3, P4, P5, P6, P7, P9, P10, P11, P13, P14, P15

[210] P1, P4, P10, P12, P15, P16

[211] P8, P11

[212] P5, P15

[213] P10

[214] P5

[215] P15

However, some researchers[216] see the evaluations themselves as the problem: 'Even if you could articulate this is the amount of risk that's societally acceptable, I think the projection of that onto an actual evaluation we have is non-obvious... I'm not sure how we would do it.'[217] One independent evaluator interviewee felt there was little incentive to develop good evaluations, with current evaluations more suited to the interests of companies than public or regulatory interests.[218]

Others connected difficulty actioning evaluations to a lack of standards in the evaluation process.[219] There is a wide range in how evaluations are conducted, presented and reported[220] with decisions on standards considered extremely challenging, especially for red teaming. Some argued for consensus on a standardised set of safety evaluations or certification for evaluators of foundation models.[221] Researchers have also called for changes to how evaluation results are reported to make it easier for policymakers to understand claims about the safety of AI systems.[222]

Others saw documentation as a key part of what makes evaluations useful.[223] Before model release, current best practices include creating:

- model cards for foundation models, which document performance biases and intended application domains[224]
- datasheets for the datasets the foundation models are trained on, which document contents, biases, intended uses, data sources and collection practices.[225]

[216] P4, P7, P9

[217] P4

[218] P15

[219] P1, P2, P3, P4, P12, P13, P14, P16

[220] Nestor Maslej, and others, 'Artificial Intelligence Index Report 2024' (2024) 169–171 https://aiindex.stanford.edu/report/ accessed 14 May 2024.

[221] P1, P4

[222] Ryan Burnell and others, 'Rethink Reporting of Evaluation Results in AI' (2023) 380 Science 136.including autonomous driving and medical diagnosis. In contexts such as these, the consequences of system failures can be devastating. It is therefore vital that researchers and policy-makers have a full understanding of the capabilities and weaknesses of AI systems so that they can make informed decisions about where these systems are safe to use and how they might be improved. Unfortunately, current approaches to AI evaluation make it exceedingly difficult to build such an understanding, for two key reasons. First, aggregate metrics make it hard to predict how a system will perform in a particular situation. Second, the instance-byinstance evaluation results that could be used to unpack these aggregate metrics are rarely made available (1

[223] P3, P7, P8, P9, P11, P12, P13

[224] Margaret Mitchell and others, 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2019) <https://doi.org/10.1145/3287560.3287596> accessed 26 June 2024.

[225] Timnit Gebru and others, 'Datasheets for Datasets' (2021) 64 Communications of the ACM 86.

Standardised documentation could help with replicability of evaluations

One interviewee compared this to other industries such as healthcare where documentation is considered to be as important as the product.[226] Standardised documentation could cover details about the model such as training data and energy consumption, which could help with replicability of evaluations.[227] There was also a desire for more procedural detail around decisions made when running evaluations, which could ensure retention of key information with staff turnover at companies.[228] However, one interviewee described companies as often reluctant to document these details over fears around liability.[229]

Evaluations can be easy to manipulate

Interviewees expressed concerns that evaluations were easy to manipulate,[230] whether intentionally or unintentionally.

Two participants were concerned that if evaluation metrics are made public, it incentivises developers to build their model to simply meet the metrics[231] rather than actually addressing the risks and harms. As one interviewee said:

> ‘Where a measure becomes a target, it ceases to become a good measure.’[232]

Reportedly, at least one foundation model developer may already be ‘gaming’ evaluations to minimise indicators of dangerous model capabilities.[233]

It is also difficult to guarantee that publicised evaluation datasets were not included in models’ training data, leading to a lack of trust in

[226] P13

[227] P6, P7, P13

[228] P3, P9, P13

[229] P13

[230] P4, P5, P8, P11, P12

[231] P7, P8

[232] P7

[233] Billy Perrigo, ‘Employees at Top AI Labs Fear Safety Is an Afterthought’ [2024] TIME https://time.com/6898961/ai-labs-safety-concerns-report/ accessed 17 May 2024.

evaluation results.[234] Data leakage can be unintentional and difficult to detect, especially when models learn from user interactions which might contain benchmark data.[235]

The lack of transparency from model developers can exacerbate a lack of trust in evaluations.[236] Some felt companies were somewhat transparent around model evaluations.[237] However, one interviewee described a recent shift away from transparency: 'Five years ago, the technical papers were so specific as to what kind of data [model developers] used… over time it became more and more limited […] Reading between the lines with all the IP litigation happening, I think most companies got smart enough to realise that what they're putting out there could be very clearly used against them.'[238] Another interviewee described competition or privacy as reasons companies may not disclose more information on model evaluations.[239]

Given the lack of standards in external evaluation, interviewees also expressed concerns that companies could potentially pick third-party evaluators they believe would give them the most favourable results[240] or could have editorial control over what is published.[241]

Lack of awareness or interest in evaluations

Evaluations are an emerging ecosystem and some actors lack awareness or interest in evaluations. Some interviewees described the challenge of making people aware of the need to evaluate and that most deployers focus on application development.[242]

One interviewee complained of the lack of interest in evaluations: 'Our whole hypothesis for building [evaluations] was that people care about

[234] P7, P11, P12, P13, P16

[235] Simone Balloccu and others, 'Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs' in Yvette Graham and Matthew Purver (eds), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2024) https://aclanthology.org/2024.eacl-long.5 accessed 26 June 2024.

[236] P9, P12, P13, P14

[237] P9, P12

[238] P3

[239] P4

[240] P3

[241] P5

[242] P1, P7, P9

Increased resources alone will not solve the challenges of evaluations

evaluation. Funnily enough, it doesn't hold that much. There's only certain people that care about evaluations and those people only care about aggregate benchmarks that they can compare against other [models] to show that their models [are] better in some way'.[243] Given this, the audience for evaluations is unclear, with some interviewees suggesting deployers aren't interested in spending money on evaluations.[244]

Summary of challenges

In summary, there are a range of challenges that currently inhibit the effectiveness of evaluations.

Challenges to the theoretical basis of evaluations are the most difficult to resolve. Current evaluations are generally focused on risks and metrics that are too abstract to meaningfully connect to the risks of real-world deployment of foundation models. Coupled with the lack of involvement of affected communities, evaluations are not adequate assessments of real-world harms or impacts on the communities most likely to be affected by their use. Addressing this disconnect in the current evaluation landscape requires rethinking how evaluations are developed, incentivising more context-specific evaluations and broadening the target of evaluation from a foundation model to a whole-system approach.

Resource constraints are also a challenge when evaluating foundation models, which can result in asymmetries regarding which companies are able to develop and conduct evaluations. Some of these challenges stem from the nature of foundation models themselves, with prompt-sensitivity and fine-tuning highlighting the ways these models are still unpredictable to developers and other evaluators. Addressing resource limitations and developing the research and evaluation ecosystem around foundation models could improve our understanding of models, which could in turn ameliorate some of the theoretical challenges.

However, the social and policy challenges along with theory challenges are substantial and won't be solved through increased resources

[243] P7
[244] P3, P7

alone. Evaluations can be manipulated and are not clearly actionable, which leaves open a question around how they fit into the broader AI accountability ecosystem if the goal is to achieve safer AI systems.

What is the role of evaluations in the broader landscape of AI governance and accountability?

Company actions as a result of evaluations

The previous chapters have established that evaluations are a potentially useful, though limited, tool for assessing the capabilities, risks and impacts of foundation models. Evaluations alone, however, do not make AI systems safe. Companies should act to address safety concerns. This can include making changes to the model, accessibility, the user interface, the training or fine-tuning data or other components that can reduce the likelihood of that risk. Our interviewees highlighted types of actions that companies might or have taken as a result of evaluations:

- **Adjusting design and training:** Companies might evaluate model performance of training checkpoints, final base models and fine-tuned models. They might then adjust, for example, changing training parameters, to improve performance.[245]
- **Adding safeguards and mitigations:** Companies might discover potential harms or risks while evaluating model checkpoints or base models. Companies might then put in place safeguards to reduce those risks, for example model fine-tuning or adjusting safety filters of model inputs and outputs.[246]
- **Delaying deployment:** Companies might choose to delay release until the model or system met certain performance or safety evaluation thresholds. Many interviewees cited OpenAI's decision to delay the release of GPT-4 to undertake red teaming and add safety guardrails as an example.[247]

[245] P6
[246] P16
[247] P7, P9, P10

- **Restricting model access:** Companies might choose to limit who could access the model. This ranged from not making the model open-source, to only sharing with approved partners and researchers, to not publicly sharing the model at all.[248]

However, companies developing these systems do not consistently take the necessary actions when evaluations do suggest cause for concern. Currently, evaluation results that demonstrate potentially significant risks rarely lead to significant slowing, much less halting or decommissioning, of foundation model research, development or deployment. OpenAI did not release GPT-2 weights, code or training data due to OpenAI's own 'concerns about malicious applications of the technology', and it has been reported that Google once slowed release of AI systems due to ethical concerns.[249] More recently, OpenAI has rapidly commercialised the much more capable GPT-3.5 and GPT-4 models through ChatGPT, and Google has reportedly sidelined ethical concerns to release its own foundation models.[250]

Several of our interviewees expressed concerns that there are currently no incentives or norms of practice from major technology companies to conduct evaluations and take meaningful actions on the results. One interviewee described a culture of wilful ignorance within some companies, claiming they actively avoided evaluations to sidestep accountability if flaws were later found.[251] This was exacerbated by a lack of clarity over who is responsible for conducting and requiring evaluations. Another participant noted that there are no strong incentives for companies to conduct or act on evaluations, in the absence of legal mandates or market pressure to make changes to applications.[252]

Another interviewee noted that many companies undertake evaluations without a clear goal or set of acceptable results and accompanying actions. The companies' actions therefore depended on executives'

[248] P12

[249] Alec Radford and others, 'Better Language Models and Their Implications' (14 February 2019) https://openai.com/research/better-language-models accessed 26 March 2024; Davey Alba and Julia Love, 'Google's Rush to Win in AI Led to Ethical Lapses, Employees Say' *Bloomberg News* (19 April 2023) https://www.bloomberg.com/news/features/2023-04-19/google-bard-ai-chatbot-raises-ethical-concerns-from-employees accessed 26 March 2024.

[250] Alba and Love (n 251).

[251] P10

[252] P1

subjective judgements post-evaluation.[253] Rather than setting objective criteria for decision-making, evaluations were instead viewed as part of a broader holistic assessment.[254]

These cases illustrate how foundation models that fail evaluations (or other ethics and safety assessments) may still be released without further checks, due to commercial pressures.

One notable exception is model disgorgement; the deletion of models or datasets as directed by regulators.[255] Disgorgement has been used several times by the US Federal Trade Commission, recently against a racially biased facial recognition algorithm used by Rite-Aid. This ruling forced the company to delete all related facial recognition datasets and models and banned them from using facial recognition for five years.[256] These same powers could be used against foundation model developers.

The contrast between companies' and regulators' actions in response to evaluation findings, such as bias and otherwise risky models, highlights the need for strong external evaluators with power to compel remedies.

Structured approaches to development and deployment decisions based on evaluations.

In response to voluntary frameworks from the UK and USA, several leading foundation model developers are now beginning to adopt a more structured approach to guide decisions about the development and deployment of foundation models. Evaluations are a core part of these approaches.

There is not yet a standardised terminology or completely agreed set of components for these evaluation-driven structured approaches to development and deployment decision-making. Anthropic has a Responsible Scaling Policy, Open AI has a Preparedness Framework

[253] P3

[254] P10

[255] Jevan Hutson and Ben Winters, 'America's Next "Stop Model!": Model Deletion' (20 September 2022) https://papers.ssrn.com/abstract=4225003> accessed 25 March 2024.

[256] Rite Aid Banned from Using AI Facial Recognition After FTC Says Retailer Deployed Technology without Reasonable Safeguards' (*Federal Trade Commission*, 19 December 2023) https://www.ftc.gov/news-events/news/press-releases/2023/12/rite-aid-banned-using-ai-facial-recognition-after-ftc-says-retailer-deployed-technology-without accessed 25 March 2024.

and Google DeepMind has a Frontier Safety Framework.[257] It is possible that many other companies are also incorporating evaluations into their decision-making but without publishing this policy commitment.

Model Evaluation and Threat Research (METR), a nonprofit that researches cutting-edge AI systems provides a useful high-level definition for these approaches. It defines these policies as those that specify 'what level of AI capabilities an AI developer is prepared to handle safely with their current protective measures, and conditions under which it would be too dangerous to continue deploying AI systems and/or scaling up AI capabilities until protective measures improve'.[258] METR is also responsible for popularising the term 'responsible scaling policies' (RSPs) and is influential on the foundation model developer commitments outlined above.

At the time of our interviews, RSPs were the most common way to refer to these emerging public commitments to structured evaluation-driven decision-making, and we will use this term within this section.

Some interviewees expressed optimism that RSPs could offer a valuable framework for more deliberate and safety-conscious decision-making in foundation model development.

Key potential benefits highlighted by interviewees:

- Providing forward-looking guidance to proactively anticipate risks and steer development in safer directions.[259]
- Increasing internal accountability within companies by setting explicit standards and risk thresholds that align decisions with safety commitments.[260]

- Establishing a foundation for further action through an iterative

[257] 'Introducing the Frontier Safety Framework' (*Google DeepMind*, 17 May 2024) https://deepmind.google/discover/blog/introducing-the-frontier-safety-framework/ accessed 17 May 2024; 'Responsible Scaling Policy, Version 1.0' (Anthropic 2023); 'Preparedness Framework (Beta)' (OpenAI 2023) https://cdn.openai.com/openai-preparedness-framework-beta.pdf accessed 17 May 2024.

[258] METR, 'Responsible Scaling Policies (RSPs)' (26 October 2023) https://metr.org/blog/2023-09-26-rsp/ accessed 14 February 2024.

[259] P2

[260] P2

process of updating RSPs based on experiences that could lead to industry standards.[261]

- Improving transparency by openly outlining risk evaluations used for development and deployment decisions.[262]

Ultimately, interviewees hoped RSPs could enable more intentional, evidence-based choices around managing AI risk in accordance with openly declared policies and thresholds.

While RSPs aim to increase transparency and accountability in AI development decisions, several interviewees raised concerns about the current standards and level of specificity. A key issue highlighted was the lack of concrete, well-defined evaluation criteria and red lines that would necessitate pausing development until stronger mitigations are in place.[263] Without clear specifications of unacceptable capabilities and robust testing methods, some experts were sceptical that existing RSPs provide sufficiently rigorous guidance.[264] OpenAI's Preparedness Framework was cited as an example lacking precise evaluation methods tied to its risk thresholds.

Beyond evaluation standards, interviewees also questioned the transparency of RSPs in practice. Concerns were raised about external oversight: how can outside parties verify whether conditions outlined in an RSP were truly triggered?[265] Some pushed for greater transparency around details like training data and model architectures to properly inform risk assessments.[266] The need for independent auditing of RSP evaluations and compliance was also emphasised, potentially through certified third-parties or regulatory standards.[267] While RSPs have the potential to improve accountability, multiple experts felt current implementations still lack the specificity and external visibility required to achieve meaningful real-world impact.

[261] P2, P16
[262] P9, P6
[263] P2
[264] P2, P12
[265] P6
[266] P9
[267] P16

Interviewees believed that governments should understand companies' RSPs, as these internal frameworks reveal the specific threat models and safety priorities driving development decisions.[268] However, they stressed that certain 'red lines' around unacceptable model capabilities cannot be left up to voluntary commitments; governments must ultimately define and mandate compliance in areas they deem too risky.[269] Some interviewees envisioned an iterative process where RSPs inform government regulations, which then get incorporated into updated RSP standards in a collaborative cycle building toward robust industry norms.[270]

Coordination challenges were highlighted as a major barrier to the effectiveness of voluntary RSP adoption.[271] Without universal buy-in, individual companies face competitive pressures that could undermine their adherence to the company's stated RSP commitments. For example, companies might expand access to models with potentially harmful biological capabilities beyond the limits stated in their RSP to increase market share.[272] This misalignment of incentives led some to argue that regulations creating an enforceable, level playing field are required.[273] Others believed RSPs demonstrating the inadequacy of self-governance could compel necessary government intervention. Several saw value in governments facilitating RSP coordination and consensus-building in the interim.[274]

Ultimately, many felt the true test of RSPs would come when a company first triggers its own 'no-go' evaluation and halts a potentially dangerous project.[275] Some interviewees were sceptical that companies would halt development and that – at most – they would delay release. Public visibility into such real-world decisions was viewed as critical for accountability and verification that policies matched actions.

While RSPs were seen as a positive step, most believed some form of government involvement, whether mandating standards, auditing

[268] P6
[269] P6
[270] P6, P2, P16
[271] P10
[272] P2
[273] P
[274] P10
[275] P6

adherence or directly regulating, would probably be needed to ensure responsible scaling across the industry. See Appendix 1 for a more in-depth discussion of the role of RSPs.

How should regulators and policymakers think about using evaluations?

As discussed in the previous chapter, most interviewees agreed that foundation model evaluations are a nascent field. Evaluations have several limitations and lack robust standards and practices.

> Policymakers or regulators should see current evaluations as a useful tool but they should be wary about basing substantial decisions solely on current evaluations without complementary evidence from other sources.

Even so, many interviewees saw potential in current and future evaluations as a tool to inform policy and regulatory decisions, although that will require further investment in regulatory expertise and third-party evaluation capacity to achieve.[276]

AI governance researchers have conducted extensive study on who should conduct evaluations and when they should be deployed in the broader context of audits.[277] The findings overwhelmingly indicate that evaluators should be external to companies creating models to ensure audits are objective and rigorous.[278] Studies of internal evaluation, AI ethics and safety efforts have uncovered numerous constraints, finding that internal actors often lack the power to effect change based on their evaluations and that companies tend to ignore negative results of

[276] P4, P16

[277] Guha and others (n 15); Inioluwa Deborah Raji and others, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) https://dl.acm.org/doi/10.1145/3351095.3372873 accessed 18 October 2023; Birhane and others, 'AI Auditing' (n 154).

[278] P1, P2, P4, P5, P6, P10, Casper and others (n 203); Daricia Wilkinson and others, 'Accountability in Algorithmic Systems: From Principles to Practice', *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery 2023) https://doi.org/10.1145/3544549.3583747 accessed 25 March 2024.

evaluations.[279]

Interviewees noted ways that evaluations could potentially be used by policymakers and regulators, in descending order of our view of their current viability:

1. As an exploratory tool, for example, to gather evidence for broader risk prioritisation or policymaking.
2. As an investigative tool for regulators scrutinising a particular model or organisation.
3. As part of a licensing or mandatory safety testing regime before a model is made available to the public or sold.

These are discussed in detail below. This is followed by discussion of potential issues faced by regulators and policymakers when engaging with and using evaluations.

Evaluations as an exploratory tool

Evaluations could be a useful exploratory tool when regulators lack a specific concern but want to broadly understand the risks, capabilities and impacts of a new AI system.[280] Evaluations could offer a structured way to uncover unexpected capabilities or harmful behaviours, instead of simply assessing the system's intended use.

This approach would not immediately trigger any action, but it would help policymakers identify emerging risks that need attention and prioritise their concerns. Evaluations would be just one piece of evidence, considered alongside theoretical analysis of potential harms and real-world monitoring of deployed models' behaviour and impacts.281

[279] Raji and others, 'Closing the AI Accountability Gap' (n 279); Inioluwa Deborah Raji and Joy Buolamwini, 'Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products', *Proceedings of the 2019 AAAI/ ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery 2019) https://doi.org/10.1145/3306618.3314244 accessed 25 March 2024; David Gray Widder and others, 'It's about Power: What Ethical Concerns Do Software Engineers Have, and What Do They (Feel They Can) Do about Them?', *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2023) https://dl.acm.org/doi/10.1145/3593013.3594012 accessed 25 March 2024.

[280] P2, P4, P9

[281] Elliot Jones, 'Keeping an Eye on AI' (*Ada Lovelace Institute* 2023) https://www.adalovelaceinstitute.org/report/keeping-an-eye-on-ai/ accessed 23 August 2023.

Evaluations as an investigative tool

Several interviewees suggested that evaluations could be a useful investigative tool when regulators have a specific concern in mind.[282] Unlike exploratory evaluations, targeted evaluations can be designed to confirm or refute suspicions about a system's potential harms. This approach was seen as more immediately promising for policy purposes than setting general evaluation-based safety requirements.

While interviewees did not think evaluations would meet the standard for legal evidence, they believed that could still be useful in guiding investigations. A poor result on targeted evaluations could act as a red flag that justifies deeper scrutiny, potentially including reviewing company internal data and processes. This could allow regulators to prioritise their limited resources towards the most high-risk models.[283]

Evaluations as a licensing tool or mandatory evaluation requirements

Making it a regulatory requirement to pass certain evaluations, with significant implications for failure (such as financial implications), will incentivise companies to train models to pass those evaluations. Some interviewees highlighted that the best evaluations were 'pointless' if they didn't result in action, and so evaluations should have consequences.[284]

For example, in the report Safe before sale, the Ada Lovelace Institute has previously argued that regulators should introduce a pre-market approval gate for foundation models. This is similar to the approach taken by the US's Food and Drug Administration (FDA) for high-risk drugs and devices.[285] This approval gate should at least entail submission of evidence to prove safety and market readiness based on internal testing and audits, third-party audits and (optional) sandboxes. Evaluations could form part of that evidence.

[282] P2, P10

[283] P14, P15

[284] P8

[285] Merlin Stein and Connor Dunlop, 'Safe before Sale' (*Ada Lovelace Institute* 2023) https://www.adalovelaceinstitute.org/report/safe-before-sale/ accessed 1 March 2024.

Interviewees highlighted that sectors such as automotive, aviation and medicines have gone through a process of developing and iterating standards and tests, and building up an evidence base on how tests translate into real-world risk and impact.[286]

Quantifying risks and setting thresholds may be more straightforward for physical systems like those in automotive and aviation sectors, where fault tolerance can be tested through methods like crash tests. However, it could prove more challenging to measure and translate the potential impacts on mental health or other societal factors into well-defined risk thresholds for AI systems.[287] There is some precedent for measuring these risks. Mental health effects are already evaluated to an extent when testing new drugs. Efforts have also been made, albeit with varying degrees of success, to assess the societal impacts of internet platforms on mental health and other factors.

Setting thresholds for risk tolerance is ultimately a political question and can come down to perceptions of risks and benefits, as much as the actual risks and benefits. Thresholds are not purely technical, but a reflection of public perceptions and balance needed to preserve public trust in regulatory institutions.[288]

Common issues for regulators and policymakers

There were several common issues raised by interviewees about the use of evaluations by regulators and policymakers in any capacity, with the issues largely becoming more problematic as evaluations move from just being (one) source of information to a hard gate on the development of foundation models and systems.

We are confident that many of these issues can and should be overcome through investment in government and third-party expertise, and the development of more rigorous and context-specific evaluations. However, policymakers and regulators should be aware of these issues when using or interpreting current evaluations and support this investment where necessary.

[286] P1, P8, P14
[287] P3, P9, P15
[288] P15

Reliability of evaluations

An issue raised by many interviewees was the problem of false positives and negatives. Imperfect evaluations could wrongly flag models as dangerous, leading to time wasted on investigations or restricting acceptably safe systems. Equally, harmful models might ‘fly under the radar’ of current evaluations.[289]

Translating outputs of available evaluation to real-world outcomes

Many interviewees stressed a fundamental disconnect between current evaluations and policymaker needs.[290] Most evaluations are designed by industry for internal use or by academics lacking the regulatory context. Even when evaluations themselves are reliable, they may not easily map onto real-world concerns. This makes them difficult for policymakers to use meaningfully.[291]

Updating and adapting to new modalities and capabilities

Interviewees emphasised that any use of evaluations in regulation must be able to adapt to the rapid progress in AI. We have seen public quantitative benchmarks become rapidly outdated by progress in AI. Any regulatory regime would need to take an adaptive approach to the evaluations used or risk them becoming obsolete.[292] Other safety-critical domains such as health already take an iterative approach, relying on guidance that can be easily updated, rather than laying down the details in legislation.[293]

Additionally, there can’t be a ‘one-size-fits-all’ approach. As models gain capabilities across different modalities (text, image, etc.), evaluations must be tailored to those modalities and their specific risks.[294]

[289] P4, P15

[290] P15

[291] P1, P10

[292] P3

[293] Clíodhna Ní Ghuidhir, ‘AI Regulation and the Imperative to Learn from History’ (*Ada Lovelace Institute*, 14 September 2023) https://www.adalovelaceinstitute.org/blog/ai-regulation-learn-from-history/> accessed 29 February 2024.

[294] P4, see also discussions of different levels of maturity of evaluations for different modalities and the difficulties of evaluating general purposes noted in the previous section.

Evaluation gaming

Interviewees worried about evaluation 'gaming' or manipulating the system. If companies know the exact evaluations used in regulation (like seeing the exam paper before the exam), they may optimise their models to pass those specific tests, without addressing the real-world risks that matter.[295] As Yarin Gal, Director of Research at UK AISI stated: 'Always assume your evaluation setting is adversarial.'[296] The 2015 Volkswagen emissions scandal (where Volkswagen admitted that since 2009, 11 million of its vehicles were equipped with software that was used to cheat on emissions tests) demonstrates how companies can be incentivised to game regulatory testing and make false claims about societal impacts to further their commercial interests.[297]

This might be addressed by keeping the details of the evaluation and related datasets confidential (to prevent companies developing workarounds), although this raises its own concerns about public accountability. This would require either sufficient in-house government expertise to develop new evaluation datasets or commission third parties to create confidential datasets.

The creation of the global network of AI safety institutes, and the EU AI Office, might provide such a source of expertise in at least those jurisdictions. It is already the case that details of UK AISI's evaluation methodology are kept confidential to prevent the risk of manipulation if revealed.[298]

A lack of in-house expertise and resource gaps

Some interviewees were concerned that regulators lacked sufficient AI expertise and that there was a significant resource gap between regulatory agencies and companies.[299] For example, the most capable

[295] P10; this a common recurring concern around setting measurable goals; this is often called 'Goodharting'.

[296] arin Gal (n 160).

[297] 'Emissions Testing Is Broken, and Other Lessons from the VW Scandal' (*Environmental and Energy Study Institute*, 2 October 2015) https://www.eesi.org/articles/view/emissions-testing-is-broken-and-other-lessons-from-the-vw-scandal accessed 28 May 2024; Deborah G Johnson and Mario Verdicchio, 'AI, Agency and Responsibility: The VW Fraud Case and Beyond' (2019) 34 AI & SOCIETY 639.

[298] AI Safety Institute (n 4).

[299] P1, P13, P15

evaluators and engineers could command higher salaries in the private sector and so would be drawn away from regulators or never join regulators in the first place. Some also expressed a concern that there simply wasn't a big pool of technical talent that industry or government could draw on, further increasing competition.[300]

Difficulty in relying on current company evaluations or third-party evaluations

Interviewees expressed doubts about regulators and policymakers relying on current company evaluations or second-party evaluations. Key concerns include:

- **Lack of standards:** Without certification or a mature market with a repository of evaluations, regulators have limited reasons to take the word of a developer or a contracted evaluator without performing their own checks.[301] In previous Ada Lovelace Institute work looking at FDA-style oversight for foundation models, we found that first-party self-assessments and second-party contracted auditing have consistently proven to be lower quality than accredited third-party or government audits.[302]

- **Financial and personal conflicts:** If evaluation firms rely on repeat business from the companies they are scrutinising, they could be incentivised to soften findings or selectively report results.[303] This undermines their use for regulatory purposes. Also, the small size of the current ecosystem means that evaluation organisations have often already directly worked with most of the companies they could be evaluating, and so may have pre-existing relationships that conflict with independence.

One interviewee suggested requiring evaluators to publicly declare assessment plans beforehand, making it harder to hide negative findings.[304]

[300] P15

[301] P1, Lara Groves and others, 'Auditing Work: Exploring the New York City Algorithmic Bias Audit Regime' (arXiv, 12 February 2024) http://arxiv.org/abs/2402.08101 accessed 19 February 2024.

[302] Stein and Dunlop (n 271) 54.

[303] P5, Groves and others (n 303).

[304] P5

Cost-burden of evaluations for developers

Interviewees raised concerns about the cost of rigorous evaluations for smaller organisations, if developers were required to undertake them themselves.[305] While the expense may not be prohibitive alongside the current high cost of model development itself, this could change as training becomes more decentralised. Smaller entities may be able to develop models but lack the resources for the in-depth evaluations that responsible deployment should require.

‘Defence in depth’ – a layered approach to risk governance

As discussed above, there are potential applications for the use of evaluations by policymakers and regulators. However, even when pre-deployment evaluations are comprehensive and transparent, they are insufficient by themselves. Evaluations and associated evaluation-driven structured decision-making around development and deployment are only one method for companies to proactively identify risks and then take steps to mitigate those risks or delay development and deployment. They should be seen as one tool alongside other AI governance tools, such as:

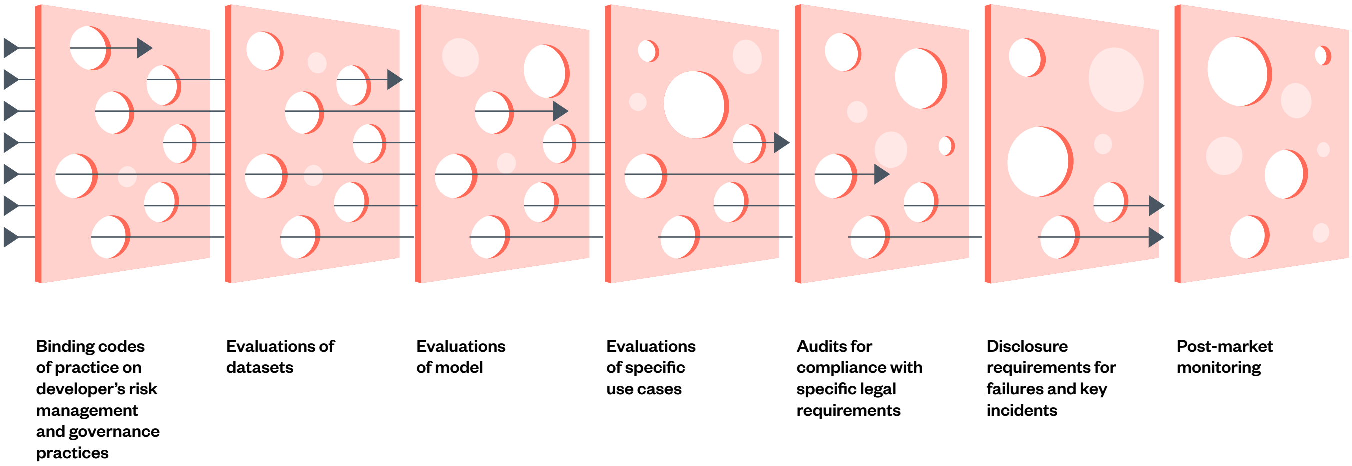
- Post-deployment monitoring and incident reporting systems, designed to identify harmful emergent behaviours in real-world settings, could provide essential data to update models and evidence to begin regulatory action.
- Well-defined liability frameworks could incentivise safer systems by attributing harms to their source.
- Regulatory-enforced roll backs of problematic deployments could also limit the impact of identified harmful applications.

The ‘Swiss Cheese model’ is a useful framework for understanding safety in complex systems. Each imperfect barrier (each slice of cheese) reduces overall risk. The holes in the layers represent potential failures at each stage. When the holes align, a failure can occur. As the holes are in different places on each ‘slice’, this minimises the likelihood of the holes aligning. The multiple layers mean potential failures from one layer

[305] P10

are likely to be 'caught' by another layer.[306] Together they minimise the chance of a catastrophic failure cascading through the whole system.

Figure 6: The 'Swiss Cheese model' applied to foundation model governance


Binding codes of practice on developer's risk management and governance practices
Evaluations of datasets
Evaluations of model
Evaluations of specific use cases
Audits for compliance with specific legal requirements
Disclosure requirements for failures and key incidents
Post-market monitoring

The development and deployment of foundation models should adopt a similar multi-layered 'defence in depth' strategy, with thorough model evaluations as one of those layers.[307]

The layers should include:

- codes of practice establishing binding practices for developers and their RSPs
- evaluations of datasets
- evaluations of model
- evaluations of specific use cases
- audits for compliance with specific legal requirements
- disclosure requirements for failures and key incidents
- post-market monitoring.

[306] Dan Hendrycks, Mantas Mazeika and Thomas Woodside, 'An Overview of Catastrophic AI Risks' (arXiv, 26 June 2023) http://arxiv.org/abs/2306.12001 accessed 4 July 2023.

[307] Shaun Ee and others, 'Adapting Cybersecurity Frameworks to Manage Frontier AI Risks: A Defense-in-Depth Approach' (Institute for AI Policy and Strategy 2023) 8 https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/6528c5c7f912f74fbd03fc34/1697170896984/Adapting+cybersecurity+frameworks+to+manage+frontier+AI+risks.pdf accessed 29 February 2024.

Governments can play an important role in improving the effectiveness of evaluations and evaluation-driven decision-making and embedding them within a broader 'defence in depth' approach.

- Governments can help to develop standards for evaluations, making it easier to identify best practices and pinpoint problematic gaps.
- Governments could also improve transparency and accountability of evaluations by requiring robust reporting mechanisms on evaluation outcomes and encouraging independent third-party audits for high-risk systems.
- Finally, only governments can legitimately establish red lines on truly unacceptable risks and capabilities from models and applications, mandating immediate interventions with clear legal backing.

No single safety safeguard is foolproof. Each has different strengths and weaknesses. But provided the holes in each layer of 'Swiss cheese' do not overlap, they can still add up to a much safer AI ecosystem.

While we have discussed many of the challenges for current evaluations, we do believe that many of these challenges are surmountable with better alignment of evaluations with regulatory needs, more external scrutiny, sufficient investment in the underlying science of evaluations and building up the third-party ecosystem. We propose these changes are worth making and the next chapter discusses in greater detail options for how evaluations can become a more effective tool in the governance toolkit.

Making evaluations a more effective part of the governance toolkit

Developing evaluations that respond to the needs and wants of regulators, policymakers and affected communities

Interviewees expressed concern about the disconnect between industry and academic evaluations and the needs of regulators and the public, highlighting the importance of developing evaluations that respond to these needs.[308] They described a landscape where rigorous evaluations are slow, under-resourced (often left to academics) and lack the clear economic or status incentives to scale quickly.[309] This raises serious questions about whether current evaluations can effectively support responsible AI governance.

Evaluations are generally either developed within companies or by academics.[310] This leads to a potential mismatch: evaluations may reflect what is easy to measure or what is of academic interest, not what regulators or the public actually need to assess risk. This is evident in the arbitrary way that benchmarks gain prominence.[311] For instance, OpenAI's choice of HellaSwag and MMLU for GPT-3 led to widespread adoption, which was driven by convenience for comparison and not necessarily the benchmarks' inherent accuracy or relevance to real-world concerns.

Misaligned evaluations may not be important in academia, but when they underpin decisions like whether to deploy or licence a model, they

[308] P4, P6, P10, P15
[309] P10
[310] P4
[311] P4

become a liability. For evaluations to be useful governance tools, they must measure outcomes that actually matter to regulators and the public. Interviewees suggested that governments should coordinate stakeholder views for evaluation developers and forecast future evaluation needs.[312]

> Interviewees saw promise in the UK AI Safety Institute's early efforts to coordinate evaluation needs and anticipate future requirements.[313]

They were optimistic about its potential to catalyse a stronger independent evaluation ecosystem. However, several stressed the importance of the institute remaining insulated from political or partisan shifts for its success.[314]

To limit the risk of evaluation gaming, policymakers and regulators may want to keep the details of some evaluation and related datasets confidential. However, this raises concerns about public accountability and would require sufficient in-house government expertise to develop these evaluation sets, rather than relying on academia or industry.

The emerging network of AI safety institutes – so far in the UK, USA, France, Canada, Singapore and Japan – and the EU AI Office, might provide such a source of expertise in at least those jurisdictions.

Building up the third-party ecosystem

As noted above, the third-party evaluation ecosystem is currently small and unstandardised, which cannot easily be addressed in the short term.[315] This creates issues for regulators and policymakers relying on them to substitute for in-house government expertise and regulatory scrutiny. In previous work, the Ada Lovelace Institute has outlined how

[312] P4, P6

[313] P2

[314] P2

[315] 'Third-Party Testing as a Key Ingredient of AI Policy' (25 March 2024) https://www.anthropic.com/news/third-party-testing accessed 30 April 2024.

the government could support the development of an AI assessment ecosystem. Those points are equally valid for supporting the foundation model evaluation ecosystem:

- create incentives for companies and third parties to evaluate foundation models
- case studies of evaluation methods in practice
- standards for evaluations
- domain or sector-specific guidance on evaluating societal risks
- skills and roles in the technology sector
- regulatory capacity building
- empowering third-party risk and impact assessors.

See our papers *AI assurance? Assessing and mitigating risks across the AI lifecycle* and *Auditing Work: Exploring the New York City algorithmic bias audit regime* for more details on each of these points.[316]

Asking questions

As discussed in the chapter on 'What is foundation model evaluation?', policymakers and regulators should ask for clarity from evaluators, developers and deployers on what is meant by evaluations in each context they are engaging with.

Questions to consider about a particular evaluation include:

- What is the target?
- Who is the audience for the evaluation?
- What would a specific evaluation result mean and what action would be linked to it?

Relatedly, policymakers should also seek clarity on what people mean when they say 'audit' and not assume 'audit' is being used distinctly or interchangeably with evaluation.

[316] Jenny Brennan, 'AI Assurance?' (*Ada Lovelace Institute* 2023) https://www.adalovelaceinstitute.org/report/risks-ai-systems/ accessed 16 August 2023; Groves and others (n 303).

Involving affected communities

The involvement of affected communities in evaluation design and decisions is crucial, as one significant limitation of foundation model evaluations is that they rarely involve groups most likely to be affected (whether positively or negatively) by foundation model deployment. It's important to recognise that the general public will be both users of foundation model applications and affected by their implementations, giving them a substantial stake in how these systems are evaluated. Yet, by not involving affected communities and the wider public, evaluations will not be able to reflect their values, needs and contexts.

The Ada Lovelace Institute has previously outlined how commercial AI companies can improve public participation across the development and deployment stages.[317] This research found that embedding public participation into projects like evaluation requires considerable resources and practitioners are concerned about extractive or exploitative public participation practices, among other challenges.[318]

One risk of exploitation may come from using voluntary participant labour to construct evaluations, with the false expectation that these evaluations will lead to the modification or removal of harmful models and applications. Participatory methods that solely seek to create a suite of evaluations may fail to avoid exploitation if decisions do not involve the public or do not reflect previously established public views on red lines and unacceptable deployment of the model.

There is more that companies and policymakers can do to involve affected communities, especially for evaluation of applications (in its deployment context):

- Policymakers and regulators, such as the newly established AI safety institutes in the UK and USA, could issue guidance for how technology companies can use public participation as part of evaluations. The establishment of a multi-stakeholder initiative around public participation methods could help create an industry norm around

[317] Lara Groves, 'Going Public: Exploring Public Participation in Commercial AI Labs' (2023) Discussion Paper https://www.adalovelaceinstitute.org/wp-content/uploads/2023/12/2023-12_ALI_Going-public_Discussion-paper.pdf accessed 19 February 2024.

[318] ibid.

the use of these methods in the AI development and deployment process, particularly in evaluation processes. Red teaming might be a particularly fruitful part of the AI lifecycle to augment with greater participatory methods.[319]

- There could also be collaborative development of standards for evaluations and evaluations themselves involving affected communities.[320] This should bring together a range of individuals, groups and organisations with expertise and background in public participation research, activism and community organising, alongside companies and governments.

Disclosure requirements and external scrutiny of company claims

Interviewees believed effective evaluations require companies to disclose more information about systems and training data than they currently do.[321] While technical and security challenges must be thoughtfully addressed, such as preventing the leaking of model weights, interviewees emphasised that these challenges cannot become an excuse for inaction or preventing external scrutiny by regulators or other third parties.[322] The Ada Lovelace Institute has previously argued that regulators should compel mandatory model and dataset documentation and disclosure for the pre-training and fine-tuning of foundation models. This should include capabilities evaluation and risk assessment within the model card for the (pre-) training stage and post-market.[323]

However, access to foundation models (beyond publicly available APIs) currently often relies on informal networks and companies' willingness to grant permission, hindering independent evaluation. This situation can affect the replicability of results and introduce bias in which researchers gain access.

[319] ibid.
[320] ibid.
[321] P14
[322] P6
[323] Stein and Dunlop (n 271) 82.

In November 2023 at the AI Safety Summit, governments and companies agreed a plan for pre- and post-deployment testing.[324] In spring 2024, press reports began to emerge suggesting that the UK AI Safety Institute has faced challenges in gaining pre-release access to the latest models of many major foundation model companies. At the time, Jack Clark, co-founder of Anthropic, described 'pre-deployment testing [as] a nice idea but very difficult to implement.'[325] Since then, Anthropic is now reported to have shared its latest model with the UK AI Safety Institute before deployment.[326] However, it is unclear whether other leading AI companies have or will provide pre-release access to their models, how far in advance they are making these models accessible and which models will be shared.

Several interviewees underscored the need for a clearer, more formalised process governing external scrutiny of such systems.[327] In an open letter, independent AI evaluators, researchers and practitioners have called for 'a legal safe harbor, protecting good-faith, public interest evaluation research provided it is conducted in accordance with well-established security vulnerability disclosure practices; and a technical safe harbor, protecting this evaluation research from account termination'.[328]

Transparency in evaluations is crucial for them to function as an accountability mechanism. Ideally, the evaluations and the outcomes of those tests should be made transparent to allow scrutiny.[329] Recent research by the Ada Lovelace Institute into auditing practice under the New York City algorithmic bias audit law finds that transparent reporting of (in this case, fairness) outcomes must be supported by robust enforcement, including penalties for companies who do not comply.[330]

[324] Prime Minister's Office, 10 Downing Street and Department for Science, Innovation and Technology, 'World Leaders, Top AI Companies Set out Plan for Safety Testing of Frontier as First Global AI Safety Summit Concludes' (*GOV.UK*, 2 November 2023) https://www.gov.uk/government/news/world-leaders-top-ai-companies-set-out-plan-for-safety-testing-of-frontier-as-first-global-ai-safety-summit-concludes accessed 25 May 2024.

[325] Cristina Criddle, Anna Gross and Madhumita Murgia, 'World's Biggest AI Tech Companies Push UK over Safety Tests' *Financial Times* (7 February 2024) https://www.ft.com/content/105ef217-9cb2-4bd2-b843-823f79256a0e accessed 1 March 2024; Vincent Manancourt, Gian Volpicelli and Mohar Chatterjee, 'Rishi Sunak Promised to Make AI Safe. Big Tech's Not Playing Ball.' POLITICO (18 April 2024) <https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/> accessed 26 April 2024.

[326] 'Introducing Claude 3.5 Sonnet' (*Anthropic*, 21 June 2024) https://www.anthropic.com/news/claude-3-5-sonnet accessed 26 June 2024; Manancourt, Volpicelli and Chatterjee (n 327).

[327] P9

[328] Shayne Longpre and others, 'A Safe Harbor for AI Evaluation and Red Teaming' (5 March 2024) http://knightcolumbia.org/blog/a-safe-harbor-for-ai-evaluation-and-red-teaming accessed 21 May 2024.

[329] P11

[330] Groves and others (n 287) 9.

To enable smooth facilitation of these processes, governments should implement measures to support an ecosystem of auditing and evaluations, including certification schemes and mandated organisational and system access. Governments cannot rely on the goodwill of AI companies, and voluntary agreements are no substitute for legal mandates to provide the access required to test foundation models, especially for pre-deployment evaluation of models and applications.

Using evaluation effectively in structured development and deployment decision-making

Leading foundation model developers are now beginning to publicly commit to a more structured approach to guide decisions about the development and deployment of foundation models. Evaluations are a core part of these approaches. However, interviewees pointed to several ways these commitments could be more effective.

Governments should encourage the development and adoption of evaluation-driven decision-making by AI companies, but they should not rely solely on voluntary commitments. As noted by interviewees, evaluations are likely to be most useful as laying the groundwork future statutory requirements for companies to comply (and be assessed for compliance) with these practices. These evaluation-based commitments should not act as a substitute to regulation and external accountability.

In the short term, through the AI Safety Institute, the UK Government should work with industry and experts to develop clearer standards for evaluations used in this structured decision-making, including specific metrics, capability limits and robust mitigation measures, including facilitating knowledge exchange between companies. This will make it clearer what risks can be effectively measured and captured under these public evaluation-driven commitments. However, as discussed in the previous section, transparency and external scrutiny will be required to establish whether these commitments are actually leading to substantive changes in decision-making around deploying potentially dangerous models.

In the medium term, the UK Government should establish its own red lines and mandatory requirements around evaluation and the

The UK Government should ensure that evaluation metrics and assumptions align with the public interest and meaningful safety

consequences of evaluation. This can address the coordination problem among AI companies by creating a level playing field through regulatory oversight and strong external evaluation measures, and so reduce the pressure on companies to prioritise market share over safety concerns.

Throughout, the UK Government should remain cautious about the concept of 'responsible scaling' as presented by AI companies, ensuring that the metrics and assumptions align with the public interest and meaningful safety.

Building a science of evaluations

Aligning incentives and targeting evaluations to regulatory priorities are important first steps. But if evaluations are to underpin regulation, a robust 'science of evaluations' would be necessary. This requires moving beyond subjective assessments towards methods grounded in verifiable results, replicable procedures and a strong theoretical underpinning.

As the UK's own AI Safety Institute states: 'Safety testing and evaluation of advanced AI is a nascent science, with virtually no established standards of best practice. The institute's evaluations are thus not comprehensive assessments of an AI system's safety, and the goal is not to designate any system as "safe".'[331] Interviewees agreed that there was a significant amount of fundamental research needed to develop the science of evaluation.[332]

As we have explored above, different approaches to evaluation are at different stages of maturity. Nevertheless, interviewees suggested that pushing forward the science of evaluation could include the below points:

- Greater emphasis on mechanistic and theory-grounded evaluations, complementing behavioural approaches to improve generalisability.
- Developing methods that are robust to variations in the prompts (whether that be rephrased text, subtly different images, reformatted code etc.), minimising the sensitivity of results to how they are prompted.

[331] AI Safety Institute (n 4).
[332] P6

- Developing a wider suite of evaluations for multimodal models, which are able to assess safety-relevant characteristics in images, videos, actions and other modalities that are currently underdeveloped relative to text-based evaluations.[333]

Apollo Research and Usman Anwar and others (2024), both set out important research questions that the field of evaluation should address as it matures.[334] They ask whether evaluations are measuring the right thing, and whether results of evaluation are trustworthy and reliable? And if not, how they can they be improved, how we can incorporate a more mechanistic understanding of models into evaluations, and how we should generalise from evaluation results, if at all?

Addressing these questions will require a concerted, long-term effort. One interviewee noted that rigorous fundamental research does not happen fast, but there is an urgency to keep pace with rapid technological advances and evolving policy requirements.[335]

Strong collaboration among stakeholders is essential. As noted above, regulators and policymakers must clearly articulate the insights they seek from evaluations. Simultaneously, the evaluation community must maintain transparency regarding existing limitations and the potential for future advancements.

Funding bodies will also need to support researchers undertaking fundamental research into evaluation science. Apollo Research has argued that both public funders (which in the UK context might include UKRI, the AI Safety Institute and the National Academies, among others) and private AI companies should provide research funding for evaluation science, with private companies standing to benefit from improved evaluation methodologies.[336] As we discuss above in 'Disclosure requirements and external scrutiny of company claims', however, there can be difficulties when independent research needs to rely on the discretion of private companies.

[333] Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (n 30) 27.

[334] 'We Need a Science of Evals' (*Apollo Research*, 22 January 2024) https://www.apolloresearch.ai/blog/we-need-a-science-of-evals accessed 1 March 2024; Anwar and others (n 40) 20–21, 54–55.

[335] P6

[336] 'We Need a Science of Evals' (n 336).

Methodology

Research questions

This paper set out to answer the following questions:

1. Compared to 'narrow' AI systems, what are the similar and unique risks that foundation models pose for people and society?
2. What is the difference between evaluation and other forms of assessment and accountability?
3. What are the proposed range of evaluation and testing approaches for addressing the risks of foundation models?
4. What are the limitations of proposed approaches to evaluation and testing?
5. What measures can be taken by policymakers to create legal/regulatory accountability of different actors in the foundation model lifecycle based on the results of evaluation and testing?

To investigate our research questions, we adopted two research methods:

- A literature review
- Expert interviews

Literature review

Our literature review was conducted between mid-December 2023 and the beginning of February 2024. We conducted a scoping review to assess the rapidly changing state of AI model evaluations and understandings of what harms and risks they should be evaluated for. We identified several existing surveys of harms taxonomies and evaluations that we synthesised into a harm taxonomy.[337] For each taxonomy, we searched for relevant papers and articles to provide an overview of harms and risks in that category in addition to the state of existing evaluations for that category.

Interviews

We conducted 16 expert interviews between December 2023 and February 2024.

We interviewed four representatives from technology companies developing foundation models, who were involved in either the development or oversight of evaluations. We interviewed a further four representatives from independent evaluation organisations, who have conducted their own evaluations of foundation models. We also spoke to eight experts from academia and civil society who have either developed or conducted evaluations themselves or have looked at the broader role of evaluations in AI governance.

Our interview questions were split into four categories to correspond with our project research questions:

- What is the meaning of evaluation and how does it relate to other kinds of assessment or accountability?
- What are existing approaches to evaluation and how are they conducted?
- What are the practical and theoretical limitations of approaches to evaluation?
- What are and should be the consequences of evaluations?

[337] Irene Solaiman and others, 'Evaluating the Social Impact of Generative AI Systems in Systems and Society' (arXiv, 12 June 2023) http://arxiv.org/abs/2306.05949 accessed 18 October 2023; Laura Weidinger and others, 'Ethical and Social Risks of Harm from Language Models' (arXiv, 8 December 2021) http://arxiv.org/abs/2112.04359 accessed 22 March 2024; Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' (Association for Computing Machinery 2021) https://doi.org/10.1145/3442188.3445922 accessed 10 March 2022.12 June 2023

Participant IDs

Participany ID	Type of organisation
P1	Government body
P2	Independent evaluator
P3	Independent evaluator
P4	Academic
P5	Independent evaluator
P6	Independent consultant
P7	Academic
P8	Academic
P9	Academic
P10	AI developer
P11	Academic
P12	AI developer
P13	AI developer
P14	Academic
P15	Independent evaluator
P16	AI developer

Partner information and acknowledgements

This work was funded by BRAID, a UK-wide programme dedicated to integrating arts and humanities research more fully into the responsible AI ecosystem, as well as bridging the divides between academic, industry, policy and regulatory work on responsible AI. BRAID is funded by the Arts and Humanities Research Council (AHRC). Funding reference: Arts and Humanities Research Council grant number AH/X007146/1.

This work was undertaken with support via UK Research and Innovation (UKRI) by the Department for Culture, Media and 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. This project was made possible by the Bridging Responsible AI Divides (BRAID) programme at the University of Edinburgh.

This paper was authored by Elliot Jones, Mahi Hardalupas and William Agnew, with substantive contributions from Andrew Strait.

We would like to thank the following people for taking the time to be interviewed for this paper:

- Rishi Bommasani, Society Lead, Stanford Center for Research on Foundation Models
- Jiahao Chen, Responsible AI LLC
- Alex Cabrera, PhD candidate, Human-Computer Interaction Institute, Carnegie Mellon University
- Melanie Mitchell, Davis Professor of Complexity, Santa Fe Institute
- Experts from Luminos.Law, Princeton University and Mohamed bin Zayed University of Artificial Intelligence
- Experts who spoke to us on background

We would also like to thank the following people for providing feedback on drafts of this paper: Leo Tarasov, Jenny Brennan, Merlin Stein and DSIT officials.

This paper also owes a lot to the prior work of Ada Lovelace Institute colleagues Jenny Brennan, Lara Groves and Connor Dunlop in investigating and clarifying many of these terms and approaches and builds directly on that work. It also benefits from comments given on previous work by Lisa Soder, Merlin Stein and Clíodhna Ní Ghuidhir.

Appendix 1: Structured approaches to development and deployment decisions driven by evaluations (‘responsible scaling policies’)

Leading foundation model developers are now beginning to adopt a more structured approach to guide decisions about the development and deployment of foundation models. Evaluations are a core part of these approaches.

There is not yet a standardised terminology or a completely agreed set of components for these evaluation-driven structured approaches to development and deployment decision-making. Anthropic has a Responsible Scaling Policy, Open AI has a Preparedness Framework and DeepMind has a Frontier Safety Framework.[338] It is possible that many other companies are also incorporating evaluations into their decision-making but without publishing this policy commitment.

Model Evaluation and Threat Research (METR), a nonprofit that researches cutting-edge AI systems provides a useful high-level definition for these approaches. They define these policies as ones that specify ‘what level of AI capabilities an AI developer is prepared to handle safely with their current protective measures, and conditions under which it would be too dangerous to continue deploying AI systems and/or scaling up AI capabilities until protective measures improve’.339 METR is also responsible for popularising the term ‘Responsible Scaling Policies’ and is influential on the foundation model developer commitments outlined above.

[338] ‘Responsible Scaling Policy, Version 1.0’ (n 259); ‘Preparedness Framework (Beta)’ (n 259); ‘Introducing the Frontier Safety Framework’ (n 259).

[339] METR (n 260).

At the time of our interviews, responsible scaling policies (RSPs) were the most common way to refer to these emerging public commitments to structured evaluation-driven decision-making, and so that's how we will refer to them for the rest of this section.

What is the promise of responsible scaling policies?

Several interviewees were optimistic that RSPs could offer a framework for more deliberate and safety-conscious decision-making within foundation model development. Experts interviewed highlighted several potential benefits:

- **Forward-looking guidance:** RSPs could provide a proactive approach, focusing on anticipating potential risks and setting actions to be taken when those risks materialise, rather than solely reacting to harms that have already occurred. This allows developers to be more intentional about steering model development and deployment in safer directions.[340]

- **Internal accountability:** RSPs necessitate setting and being explicit about internal standards and defining acceptable levels of risk.[341] This can increase accountability within organisations, as staff are more able to make and question decisions with reference to an agreed and openly debated framework, ensuring decision-making about development and deployment aligns with stated safety commitments.

- **Foundation for further action:** The development of RSPs will be iterative, as companies update their risk thresholds and evaluations in response: to their own experiences with evaluations, development and deployment; the experiences of other companies; government and regulatory action; and feedback from outside experts.[342] Some interviewees hoped that this process of setting out and iterating RSPs and their related evaluations would also lead to the development of industry standards for dangerous capability evaluations.[343] They then believed that this would lead to a 'race to the top' in terms of standards, with all foundation model developers feeling they needed to set out

[340] P2
[341] P2
[342] P6, P2
[343] P2

risk thresholds and test against those thresholds.[344] It would also act as a basis for governments mandating standards for development and deployment decisions.[345]

- **Increased transparency:** By explicitly outlining risk thresholds and the evaluations used to measure them, interviewees hoped RSPs could improve transparency about development and deployment decision-making.[346] Governments and the public could get a better understanding of company decision-making, and what evidence they will use to make and justify those decisions. For example, the choice of risks and subsequent evaluations indicates which threat models companies are most concerned about.[347]

Do responsible scaling policies currently improve transparency and accountability in development and deployment decision-making?

While interviewees highlighted some benefits of responsible scaling policies, there were also concerns about the standards and specificity of evidence and whether we were in a position to produce meaningful information to inform decision-making from evaluations. Others went further and questioned how transparent current responsible scaling policies were in practice.

Several interviewees raised concerns about the lack of concrete standards for determining what constitutes reliable evidence in RSP evaluations. One expert emphasised the need for companies to move beyond general principles and define the below:[348]

- Clear red lines: Companies should specify precise evaluations that, if failed, would necessitate a pause on further model development until stronger mitigations are in place.

- Unacceptable capabilities: Alongside these ‘tripwire’ tests, companies

[344] P16
[345] P2
[346] P9
[347] P6
[348] P2

should explicitly state which capabilities their models must not demonstrate during evaluation. Failures on these metrics would signal fundamental safety concerns.

- Robust mitigations: These specific evaluations and capability limits should be supported by rigorous information security practices and other safety measures to actively prevent undesirable outputs or behaviours.

Interviewees expressed varied levels of scepticism about the existing RSPs, noting a lack of specificity hindering real-world impact.[349] OpenAI's Preparedness Framework, while providing broader behavioural descriptions, lacks clear evaluation methods. This raised concerns that their risk thresholds might be too lenient for the capabilities envisioned. Another expert stressed that without the development of evidence-based, empirically tested assessment standards, current RSPs lack sufficient rigour to provide meaningful guidance.[350]

Furthermore, while decision-making transparency was raised as a potential benefit of RSPs, some interviewees did have concerns about the current level of transparency in practice. Primarily, the question of external oversight: how can policymakers and other stakeholders verify whether a company has actually triggered a condition outlined in its RSP?[351] Citing OpenAI's reported delay of GPT-4, one interviewee argued that such decisions, even when made seemingly in the interest of safety, require far greater transparency and opportunities for external input.[352]

Interviewees also stressed the need for regulators to have access to details about training data and model architecture.[353] This level of detail would not only improve accountability but also help inform risk mitigation strategies. Knowing whether concerning outputs were the result of memorised training data, emergent capabilities or in-context learning, for example, would be important information in determining whether and in what form any new restrictions or safety measures are required.

The ability to independently scrutinise and audit RSPs was also raised

[349] P2
[350] P12
[351] P6
[352] P9
[353] P9

as a concern. One interviewee questioned the maturity of the third-party ecosystem for assessing the evaluations performed by companies and subsequent follow-through in compliance by companies with their own stated policies.[354] They thought that a regulator or other public sector body should provide certification or validation of third-party auditors, or at least provide some standards for best practice.

Should governments rely on voluntary commitments to responsible scaling from companies?

Interviewees believed that it is important for government agencies to be familiar with and understand companies' internal policies, such as RSPs.[355] They believe that doing so will help governments understand the strengths of RSPs and their evaluation-based decision-making, along with the specific threat models those evaluations target.

While acknowledging the potential value of companies establishing RSPs, interviewees also stressed that certain elements should not be left up to voluntary policies by individual companies.[356] If there are red lines, that is, model capabilities deemed categorically unacceptable, governments have the legitimacy to determine when those lines are crossed and mandate clear restrictions. Ultimately, governments need to decide where they establish their own absolute limits and require mandatory compliance.

As mentioned above, interviewees believed this would lead to an iterative process of RSPs informing government decision-making and red lines and this, in turn, informing updates to RSPs.[357] Interviewees expressed hope that the process of companies setting out and iterating on RSPs, along with their related evaluations, could lay the groundwork for industry standards focused on dangerous capability assessments. They see this as a potential springboard for more robust government action in the future.

This iterative process could involve companies sharing successes

[354] P16
[355] P6
[356] P6
[357] P6, P2

and failures in a constructive setting, potentially facilitating more widespread agreement on best practices.[358] Ideally, one interviewee hoped, a collaborative environment involving policymakers would guide this knowledge exchange, creating the foundation for either voluntary industry consensus or more concrete regulatory frameworks.

Several interviewees shared the aspiration that RSPs could lead to a ‘race to the top’ regarding safety standards.[359] They envision foundation model developers collectively adopting rigorous risk thresholds and testing themselves against those standards.

Relatedly, another interviewee believed that there was an argument for governments helping those who do want to adopt RSPs to coordinate on reasonable standards for policies.[360] More generally, they saw the value of RSPs and similar frameworks as helping to create consensus among companies and other stakeholders on, for example, what risks do and do not exist.[361]

The importance and difficulty of coordination

Some interviewees emphasised the importance of coordination between companies in the effectiveness of RSPs. They noted that the importance of RSPs and deployment decisions partly depends on what others have committed to.[362] If a company commits to not releasing a model with a given dangerous capability, but knows others will release a model with those capabilities, then the company’s principled stance has a much lower counterfactual impact. Interviewees agreed that, ideally, companies should still consider the impact of their actions unilaterally; however, in practice, clear coordinated pre-commitments about red lines and what they won’t build or deploy could be effective in ensuring that RSPs are counterfactually effective.

However, others expressed scepticism about the effectiveness of RSPs without regulation or enforced government coordination, due to the

[358] P16
[359] P2
[360] P10
[361] P10
[362] P10

economic realities of competition. Companies aiming to limit deployment of risky systems or enhance internal security standards face potentially losing market share to less cautious competitors.

Thus, one interviewee believed that while companies might sign-up to industry standards, in practice, economic incentives would pressure them not to follow-through on their stated commitments.[363] They still believed it was valuable for companies to adopt RSPs; those RSPs are likely to fail but will make it evident to governments and the public that self-regulation is not effective and government intervention is required to address this coordination problem. They believed that if regulatory oversight and strong external evaluation measures apply equally to all companies, this would level the playing field and strengthen incentives for adherence to safety standards.

Other interviewees also argued in favour of government intervention. They noted that companies have different priorities and incentives to governments and the public, and so while it was promising and encouraging that companies had adopted these policies, they did not want society to be fully dependent on them voluntarily making important safety-related decisions.[364] As long as these policies are voluntary actions, they would be subject to change. Incentives and company personnel will change over time, and so enforcement and interpretation of the policies might be erratic and not aligned to the broader interests of society.[365]

One interviewee noted that governments have already indirectly influenced companies towards responsible scaling practices.[366] Existing regulations that establish compute thresholds, such as the Biden Administration's Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence and the EU AI Act, have essentially forced companies to implement some form of risk assessment and evaluation to remain compliant.

However, they were also concerned that these thresholds were poor proxies for risk and a crude substitute for comprehensive, risk-based

[363] P2
[364] P6
[365] P6
[366] P12

policies. They believed that more nuanced government strategies focusing on early indicators of a broad spectrum of risks are essential. While progress is being made, they believed there was a clear need for government policies that directly target specific safety concerns rather than relying solely on proxy metrics.

The acid test

Ultimately, many interviewees agreed that a critical test of RSP effectiveness will lie in their real-world application. When (and if) a company's evaluation triggers a 'no-go' decision, that will be the true test of whether RSPs can function as intended to halt the development of potentially dangerous models, but we are yet to (publicly) see such a test.[367]

Is 'responsible scaling' even a meaningful concept?

Finally, one interviewee was sceptical that the idea of 'responsible scaling' as presented by companies was at all meaningful.[368]

They did not believe that the technical mechanisms of evaluation were necessarily a problem, and they believed that you could build reasonable evaluations under the definition given for responsible safety. But they believed the metrics for responsible scaling and the assumptions underlying the choice of metrics was the problem.[369]

They viewed responsible scaling as representing a corporate vision of what AI safety should be, defined to benefit those corporations instead of benefiting and protecting the public. They believed that this definition would then be a distraction to governments, who would follow that definition of safety, on the assumption it was a fair understanding of safety. The interviewee didn't not believe current RSPs match traditional understandings of safety.[370]

[367] P6
[368] P15
[369] P15
[370] P15

Appendix 2: Current state of foundation model evaluations

Performance evaluations

AI systems are trained to optimise for certain objectives. There are many training objectives, such as the average difference between AI-generated and real images for text-to-image models,[371] or the likelihood of generating the next word in a sentence for LLMs.[372] These objectives are typically mathematically differentiable and fast to compute, limiting the types of objectives that can be used. These objectives typically encode that the AI system should reproduce and generalise from data in its training set. AI system performance is often characterised in terms of these training objectives; that is, higher performing AI systems are those that successfully reproduce and generalise the patterns in their training data.

A wide range of tests exist for evaluating model behaviour on specific tasks, including answering yes/no questions about texts,[373] answering commonsense questions,[374] and answering multiple choice science questions.[375] These evaluations take the form of lists of test questions or examples, and the AI system is rated by the fraction of questions it answers correctly.

[371] Robin Rombach and others, 'High-Resolution Image Synthesis with Latent Diffusion Models', *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (2022) https://ieeexplore.ieee.org/document/9878449 accessed 26 June 2024.

[372] Matthew Burtell and Helen Toner, 'The Surprising Power of Next Word Prediction: Large Language Models Explained, Part 1' (*Center for Security and Emerging Technology*, 8 March 2024) https://cset.georgetown.edu/article/the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/ accessed 26 June 2024.

[373] Christopher Clark and others, 'BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions' in Jill Burstein, Christy Doran and Thamar Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Association for Computational Linguistics 2019) https://aclanthology.org/N19-1300 accessed 26 June 2024.

[374] Rowan Zellers and others, 'HellaSwag: Can a Machine Really Finish Your Sentence?' in Anna Korhonen, David Traum and Lluís Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) https://aclanthology.org/P19-1472 accessed 26 June 2024.

[375] Johannes Welbl, Nelson F Liu and Matt Gardner, 'Crowdsourcing Multiple Choice Science Questions' in Leon Derczynski and others (eds), *Proceedings of the 3rd Workshop on Noisy User-generated Text* (Association for Computational Linguistics 2017) https://aclanthology.org/W17-4413 accessed 26 June 2024.

Human raters and annotators are frequently used to provide performance evaluations better aligned with human perception and preferences. Image generation model performance may be evaluated by measuring how often humans prefer AI-generated images to real images.[376] Similarly, human raters can compare AI-generated text to human-generated text to evaluate language-generation models.[377]

Bias, stereotypes and representational harms

Through their training data, AI systems and datasets can encode and even amplify societal biases and stereotypes.

AI systems have repeatedly been shown to have worse performance for women than for men and for people of colour than for white people.[378] Intersectionally marginalised people in particular, for example Black women, are less represented in datasets used to train AI systems. AI datasets are frequently biased against poorer countries, containing significantly less data on these populations.[379] Consequently, AI systems have worse performance for these groups and often perpetuate stereotypes about them.[380]

Similarly, facial recognition models perform worse for women than for men, worse for Black people than for white people, and worst of all for Black women.[381] This disparate performance can have serious impacts, including wrongful arrests and medical AI making incorrect decisions.[382]

[376] Rombach and others (n 373).

[377] Yupeng Chang and others, 'A Survey on Evaluation of Large Language Models' (2024) 15 ACM Transactions on Intelligent Systems and Technology 1.

[378] Su Lin Blodgett and others, 'Language (Technology) Is Power: A Critical Survey of "Bias" in NLP' (arXiv, 29 May 2020) http://arxiv.org/abs/2005.14050 accessed 22 March 2024; Andrew Hundt and others, 'Robots Enact Malignant Stereotypes', *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery 2022) https://dl.acm.org/doi/10.1145/3531146.3533138 accessed 22 March 2024.

[379] Dodge and others (n 11); Shreya Shankar and others, 'No Classification without Representation: Assessing Geodiversity Issues in Open Data Sets for the Developing World', *NIPS 2017 workshop: Machine Learning for the Developing World* (2017).

[380] Rida Qadri and others, 'AI's Regimes of Representation: A Community-Centered Study of Text-to-Image Models in South Asia', *2023 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2023) https://dl.acm.org/doi/10.1145/3593013.3594016 accessed 9 February 2024.

[381] Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', *Conference on Fairness, Accountability and Transparency* (PMLR 2018) http://proceedings.mlr.press/v81/buolamwini18a.html accessed 6 April 2021.

[382] Peter A Noseworthy and others, 'Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis' (2020) 13 Circulation. Arrhythmia and electrophysiology e007988; Thaddeus L Johnson and others, 'Facial Recognition Systems in Policing and Racial Disparities in Arrests' (2022) 39 Government Information Quarterly 101753.

Furthermore, datasets often contain toxic stereotypes, such as popular image and text datasets containing many pornographic images of women.[383] AI systems can often produce hateful, offensive, toxic or otherwise undesirable content. Microsoft's Tay chatbot infamously spewed racist and antisemitic tweets.[384] Image foundation models have been used to generate deepfake nonconsensual pornography using photographs of people.[385]

In addition, AI system outputs can be biased towards the cultures and worldviews reflected in their training data. For example, Midjourney, a popular AI image generator, encodes western norms of smiling in generated selfies and images of people from different cultures,[386] reflecting a broader underlying bias.

These harms can be evaluated for across several kinds of outputs. These include assessing for bias in simple text questions and social media post evaluations, images and multimodal models.[387] A variety of evaluations exist to assess the capacity of models to generate and process hate speech in text modalities.[388] However, what is considered offensive or hateful is often contextual, political and constantly evolving. This complicates the task of generating a set of static, universal evaluations for sensitive and offensive content.

[383] Abeba Birhane, Vinay Uday Prabhu and Emmanuel Kahembwe, 'Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes' (arXiv, 5 October 2021) http://arxiv.org/abs/2110.01963 accessed 22 March 2024.

[384] Matthew Handelman, 'Artificial Antisemitism: Critical Theory in the Age of Datafication' (2022) 48 Critical Inquiry 286.

[385] Emanuel Maiberg, 'Inside the AI Porn Marketplace Where Everything and Everyone Is for Sale' *404 Media* (22 August 2023) https://www.404media.co/inside-the-ai-porn-marketplace-where-everything-and-everyone-is-for-sale/ accessed 22 March 2024.

[386] jenka, 'AI and the American Smile' (*Medium*, 28 March 2023) https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf accessed 26 June 2024.

[387] Moin Nadeem, Anna Bethke and Siva Reddy, 'StereoSet: Measuring Stereotypical Bias in Pretrained Language Models' in Chengqing Zong and others (eds), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Association for Computational Linguistics 2021) https://aclanthology.org/2021.acl-long.416 accessed 26 June 2024; Dora Zhao, Angelina Wang and Olga Russakovsky, 'Understanding and Evaluating Racial Biases in Image Captioning', *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021) https://ieeexplore.ieee.org/document/9711470 accessed 22 March 2024; Saloni Dash, Vineeth N Balasubramanian and Amit Sharma, 'Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals', *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (2022) https://ieeexplore.ieee.org/document/9706958 accessed 22 March 2024; Parrish and others (n 53).

[388] Samuel Gehman and others, 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models' in Trevor Cohn, Yulan He and Yang Liu (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020* (Association for Computational Linguistics 2020) https://aclanthology.org/2020.findings-emnlp.301 accessed 26 June 2024; Paul Röttger and others, 'Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models' in Kanika Narang and others (eds), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)* (Association for Computational Linguistics 2022) https://aclanthology.org/2022.woah-1.15 accessed 26 June 2024.

Evaluations can occur at every stage of AI data collection, design and deployment. Most existing bias evaluations focus on the datasets AI is trained on, or the AI before deployment. These evaluations can show general biases of AI datasets and models but are not sufficient to understand biases that occur in specific applications and contexts. Evaluations are application focused, leaving a huge gap between existing evaluations and understanding of contexts in which harms could occur, a gap identified by several of our interviewees.[389]

Evaluations for bias and disparate performance in applications exist for some types of facial analysis, simple text questions, chest X-ray diagnosis and diagnosis of eye diseases, among others.[390] However, these are countless applications where no such disparate performance evaluations exist. Evaluating disparate performance faces the same difficulties as evaluating bias, that is, how to decide the groups on which performance should be evaluated.

Furthermore, deciding which groups of people to choose, to assess a system for bias risks is a challenging question. Bias is often highly contextual, and evaluators may not have the knowledge necessary to understand which groups are at risk from a particular AI system application.[391] Bias is also intersectional, and finding sufficient participants from intersectionally marginalised groups can be challenging.

Finally, groups experiencing bias from AI models are also often marginalised in social and political settings. These groups may be poorly represented in decision-making processes to decide what and how to evaluate, leading to evaluations that do not attempt to assess bias against them.

[389] P4, P11, P13, P15

[390] Kimmo Kärkkäinen and Jungseock Joo, 'FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation', *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021) https://ieeexplore.ieee.org/document/9423296 accessed 26 June 2024; Parrish and others (n 53); Muhammad Osama Khan and others, 'How Fair Are Medical Imaging Foundation Models?', *Proceedings of the 3rd Machine Learning for Health Symposium* (PMLR 2023) https://proceedings.mlr.press/v225/khan23a.html accessed 22 March 2024; Yan Luo and others, 'FairVision: Equitable Deep Learning for Eye Disease Screening via Fair Identity Scaling' (arXiv, 12 April 2024) http://arxiv.org/abs/2310.02492 accessed 26 June 2024.

[391] Alex Hanna and others, 'Towards a Critical Race Methodology in Algorithmic Fairness', *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM 2020) https://dl.acm.org/doi/10.1145/3351095.3372826 accessed 22 March 2024.

Privacy, data protection and intellectual property

Foundation models pose a variety of privacy risks.

LLMs have been repeatedly shown to regurgitate sensitive information from training, including phone numbers and email addresses.[392] LLMs also have potential to automate some aspects of doxing or intensive stalking, combined with release of personal information.[393] Additionally, foundation models are trained on vast quantities of copyrighted data produced by artists, writers and journalists. Foundation models are then able to, for example, generate art with the same content and style as human artists, causing economic and other harms.[394]

Some evaluations have been proposed to benchmark capacity of LLMs to leak personal information in different settings.[395] However, these evaluations are challenging because of the ease of 'jailbreaking' or hacking LLMs. 'Jailbreaking' LLMs is the practice of finding specific text inputs that cause the LLM to ignore previous safety instructions and training. For example, to evade training against outputting personally identifiable information (PII), an attacker might first instruct an LLM that it is a detective in a fictional story, bypassing restrictions about outputting PII of real people.

While seemingly ridiculous, similar jailbreaks are easy to generate and effective against even the most advanced production LLMs, causing existing evaluations to not be indicative of overall security.[396] There has been little published work on systematically assessing the presence of copyrighted content. In addition, there are few methods of assessing comprehensive categories of PII in modalities such as images, audio or video.

[392] Chen and others (n 124); Nasr and others (n 10).

[393] Robin Staab and others, 'Beyond Memorization: Violating Privacy Via Inference with Large Language Models' (arXiv, 11 October 2023) http://arxiv.org/abs/2310.07298 accessed 22 March 2024.11 October 2023

[394] Dodge and others (n 11); Alexandra Alter and Elizabeth A Harris, 'Franzen, Grisham and Other Prominent Authors Sue OpenAI' The New York Times (20 September 2023) https://www.nytimes.com/2023/09/20/books/authors-openai-lawsuit-chatgpt-copyright.html accessed 22 March 2024; Abeba Birhane and others, 'Into the LAIONs Den: Investigating Hate in Multimodal Datasets' (arXiv, 6 November 2023) http://arxiv.org/abs/2311.03449 accessed 9 November 2023; Michael M Grynbaum and Ryan Mac, 'The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work' *The New York Times* (27 December 2023) https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html accessed 22 March 2024.

[395] Chen and others (n 124)

[396] Patrick Chao and others, 'Jailbreaking Black Box Large Language Models in Twenty Queries' (arXiv, 13 October 2023) http://arxiv.org/abs/2310.08419 accessed 22 March 2024.13 October 2023

Environmental costs

Foundation models have significant energy and resource costs. Training foundation models requires cutting-edge hardware, and in particular the most advanced graphics processing units (GPUs). As a result, outdated and worn-out hardware components generate toxic electronic waste.[397]

Data centres use significant amounts of energy. According to International Energy Agency, data centres, cryptocurrencies and AI together consumed ~460 TWh of electricity worldwide in 2022, almost 2% of total global electricity demand.[398] In their base case, this is forecast to nearly double to 800TWh by 2026. The UK National Grid estimates data centres could represent up to 6% of UK electricity demand by 2030 compared with around 1% in 2023, although there is still considerable uncertainty about future data centre energy demands.[399] Data centres use large quantities of water for cooling. Total water usage by data centres is rapidly growing, with mid-sized data centres using the same amount of water as 100,000 homes.[400] AI training workloads likely play a significant factor in this growth, with training of even mid-sized AI models emitting nearly 25 tonnes of CO2,[401] the same amount as six mid-sized cars in a year.[402]

Some benchmarks exist for evaluating AI models' energy use – but these are in controlled laboratory settings.[403] While informative, these evaluations are missing real-world context. Information includes rates at which different models and features are used; energy and water usage of both devices and data centres where models are deployed; and local availability of clean electricity and water. In addition, there is also a lack of analysis of the impact on local energy economics in places where foundation models are developed and used.

[397] Kate Crawford and Vladan Joler, 'Anatomy of an AI System' (*Anatomy of an AI System*, 2018) http://www.anatomyof.ai accessed 22 March 2024; Emma Strubell, Ananya Ganesh and Andrew McCallum, 'Energy and Policy Considerations for Deep Learning in NLP' (arXiv, 5 June 2019) http://arxiv.org/abs/1906.02243 accessed 22 March 2024; Cornelis P. Baldé and others, 'Global E-Waste Monitor 2024' (International Telecommunication Union (ITU) and United Nations Institute for Training and Research (UNITAR) 2024).

[398] IEA, 'Electricity 2024 - Analysis and Forecast to 2026' (IEA 2024) https://www.iea.org/reports/electricity-2024.

[399] 'Future Energy Systems 2023' (National Grid Electricity System Operator 2023) 70 https://www.nationalgrideso.com/document/283101/download accessed 27 June 2024.

[400] Coco Zhang, Jan Frederik Slijkerman and Diederik Stadig, 'Growth in Water Consumption of Data Centres Needs More Attention' (*ING Think*) https://think.ing.com/articles/data-centres-growth-in-water-consumption-needs-more-attention/ accessed 16 May 2024.

[401] Luccioni, Viguier and Ligozat (n 50).

[402] OAR US EPA, 'Greenhouse Gas Emissions from a Typical Passenger Vehicle' (12 January 2016) https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle accessed 16 May 2024.

[403] Peng and others (n 137).

Data centres, including those used to train AI systems, can significantly shift local energy economics, contributing to coal plants remaining online.[404] Understanding the environmental impacts of resource extraction and manufacturing to build chips, and of data centres used to train and run foundation models, is notoriously difficult, with even major technology companies struggling to audit their supply chains.[405]

Disinformation, misinformation and overreliance

Foundation models have a high risk of amplifying misinformation, or false or inaccurate information, through creation of fake stories, images and even videos or audio recordings.

AI systems have been used to spread disinformation, false information intended to mislead, through, for example, impersonating political leaders including Joe Biden and former Indonesian President Suharto.[406] They have also been used to inflame conflicts including the war on Gaza.[407] However, it is important to note that algorithmic misinformation is not new or unique to foundation models, and has been amplified by search and recommender systems for decades.[408]

[404] Josh Saul and Saijel Kishan, 'AI Needs So Much Power That Old Coal Plants Are Sticking Around' *Bloomberg.com* (25 January 2024) https://www.bloomberg.com/news/articles/2024-01-25/ai-needs-so-much-power-that-old-coal-plants-are-sticking-around accessed 22 March 2024.

[405] Crawford and Joler (n 399); Brendan Sinclair, 'Microsoft's Concerning Conflict Minerals Disclosure Reflects Industry-Wide Slippage' *GamesIndustry.biz* (2 August 2023) https://www.gamesindustry.biz/microsofts-concerning-conflict-minerals-disclosure-reflects-industry-wide-slippage accessed 22 March 2024.

[406] Allegra Rosenberg, 'AI-Generated Audio of Joe Biden and Donald Trump Trashtalking While Gaming Is Taking over TikTok' *Business Insider* (1 March 2023) https://www.businessinsider.com/voice-ai-audio-joe-biden-donald-trump-tiktok-2023-3 accessed 22 March 2024; Arie Firdaus, 'Fake Suharto Video Fuels Debate on AI Use in Indonesian Election Campaign' *Benar News* https://www.benarnews.org/english/news/indonesian/suharto-deepfake-used-in-election-campaign-01122024135217.html accessed 22 March 2024; Krzysztof Węcel and others, 'Artificial Intelligence—Friend or Foe in Fake News Campaigns' (2023) 9 Economics and Business Review https://journals.ue.poznan.pl/ebr/article/view/736 accessed 22 March 2024.

[407] David Klepper, 'Fake Babies, Real Horror: Deepfakes from the Gaza War Increase Fears about AI's Power to Mislead' *AP News* (28 November 2023) https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47 accessed 22 March 2024.

[408] Danaë Metaxa and Nicolás Torres-Echeverry, 'Google's Role in Spreading Fake News and Misinformation' (31 October 2017) https://papers.ssrn.com/abstract=3062984 accessed 22 March 2024; Fabiana Zollo and Walter Quattrociocchi, 'Misinformation Spreading on Facebook' in Sune Lehmann and Yong-Yeol Ahn (eds), *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks* (Springer International Publishing 2018) https://doi.org/10.1007/978-3-319-77332-2_10 accessed 22 March 2024.

A few evaluations exist to measure the ability of LLMs to generate misinformation about specific topics, or the frequency and the types of incorrect or hallucinated information foundation models generate.[409] However, these evaluations do not consider real-world threats, where disinformation actors can use 'jailbreaks' or other techniques to hack foundation models. In addition, these evaluations assess misinformation generation in laboratory contexts, but do not assess the real-world impacts of this misinformation.

There are many evaluations to measure the ability of LLMs to detect LLM-generated text. However, automated detection of LLM generated text has proven difficult and prone to bias against non-native speakers.[410]

Foundation models can also cause harm through overreliance, especially on critical tasks or tasks where they have poor performance. Two US lawyers were recently fined for using ChatGPT to create legal filings that included hallucinated, non-existent legal cases.[411] AI-generated books on mushroom foraging contain potentially lethal misinformation on species identification.[412] No evaluations exist to measure this risk, which is both task-dependent and requires human input to measure accurately.[413]

Inequality, marginalisation and violence

When deployed in the real world, where complex social and political systems exist, systems that are biased, reproduce toxic stereotypes and

[409] Vipula Rawte, Amit Sheth and Amitava Das, 'A Survey of Hallucination in Large Foundation Models' (arXiv, 11 September 2023) http://arxiv.org/abs/2309.05922 accessed 22 March 2024; Jiawei Zhou and others, 'Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions', *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (ACM 2023) https://dl.acm.org/doi/10.1145/3544548.3581318 accessed 22 March 2024.

[410] Weixin Liang and others, 'GPT Detectors Are Biased against Non-Native English Writers' (2023) 4 Patterns https://www.cell.com/patterns/abstract/S2666-3899(23)00130-7 accessed 26 June 2024; Yafu Li and others, 'MAGE: Machine-Generated Text Detection in the Wild' (arXiv, 21 May 2024) http://arxiv.org/abs/2305.13242 accessed 26 June 2024; Canyu Chen and Kai Shu, 'Can LLM-Generated Misinformation Be Detected?' (arXiv, 16 March 2024) http://arxiv.org/abs/2309.13788 accessed 22 March 2024.

[411] Dan Milmo, 'Two US Lawyers Fined for Submitting Fake Court Citations from ChatGPT' *The Guardian* (23 June 2023) https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt accessed 16 May 2024.

[412] Dan Milmo, 'Mushroom Pickers Urged to Avoid Foraging Books on Amazon That Appear to Be Written by AI' *The Guardian* (1 September 2023) https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai accessed 16 May 2024.

[413] Raja Parasuraman and Victor Riley, 'Humans and Automation: Use, Misuse, Disuse, Abuse' (1997) 39 Human Factors 230; Weidinger and others, 'Ethical and Social Risks of Harm from Language Models' (n 339).

underrepresent already marginalised communities can create severe, compounding harms.

Repeated censorship can lead to community erasure, rendering marginalised communities nearly invisible online and degrading the ability of members of those communities to find each other online and thereby exist, at least in online form.[414] Promotion of, or failure to moderate hateful and toxic content can incite real-world violence, for example, against the Rohingya in Myanmar.[415]

Some of these harms often have roots in bias, representation and stereotype harms – and therefore evaluations of those harms are informative. On the other hand, harms relating to inequality, marginalisation and violence result from complex interactions with social and political systems that are extremely challenging to model or evaluate.

Weapons development

Foundation model applications for weapons, mass surveillance and militarisation are under development.

These uses have the potential to lead to more powerful and wider-scale cyber attacks and autonomous weapons.[416] It is also theorised that LLMs could facilitate the development of biological weapons by providing access to specialised biological and chemical knowledge.[417]

Current evaluations of these capabilities are conducted via red teaming or simulating users attempting to use foundation models for weapons development.[418] Existing evaluations have found minimal marginal

[414] Oliver L Haimson and others, 'Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas' (2021) 5 Proceedings of the ACM on Human-Computer Interaction 466:1.

[415] 'Myanmar: Facebook's Systems Promoted Violence against Rohingya; Meta Owes Reparations – New Report' (Amnesty International 2022) https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/ accessed 22 March 2024.

[416] Weidinger and others, 'Sociotechnical Safety Evaluation of Generative AI Systems' (n 30); Julian Hazell, 'Spear Phishing With Large Language Models' (arXiv, 22 December 2023) http://arxiv.org/abs/2305.06972 accessed 22 March 2024.

[417] Igor Rubinic and others, 'Artificial Intelligence in Clinical Pharmacology: A Case Study and Scoping Review of Large Language Models and Bioweapon Potential' (2024) 90 British Journal of Clinical Pharmacology 620.

[418] OpenAI, 'GPT-4 System Card' (2023) https://cdn.openai.com/papers/gpt-4-system-card.pdf accessed 22 March 2024.

risks.[419] In addition, LLMs remain vulnerable to 'jailbreaks' and other creative workarounds of safety measures, limiting the generality of any particular evaluation.

Surveillance and censorship

AI models are used for a wide array of surveillance applications.

Facial recognition and analysis websites collect massive databases of images and associated metadata, allowing anyone to identify people from face images alone and uncover potentially sensitive information, including names, addresses, emails and phone numbers.[420] Facial recognition has been adopted by law enforcement in many jurisdictions, leading to wrongful arrests.[421] Facial recognition has been widely deployed against protestors, including in the USA, Russia, China and India. It has enabled authorities to identify, track and potentially punish protestors.[422]

Natural language processing AI models are also used in many surveillance applications. Voice transcription AI models have seen wide adoption for surveilling prisoners' phone calls, with prison authorities using this technology to identify calls they could use to defend themselves against lawsuits over prison uncleanliness.[423] Employers use AI models to surveil employees' emails, messages and social media posts, and intensively surveil and monitor the calls of call-centre workers.[424]

More advanced AI systems have the potential to intensify surveillance capacities, allowing for more complex analysis of communications and enabling surveillance to fuse multiple modalities of data, including text, images, audio and location data. There are no comprehensive evaluations for the capacity of AI systems to intensify surveillance.

[419] Sayash Kapoor and others, 'On the Societal Impact of Open Foundation Models' (arXiv, 27 February 2024) http://arxiv.org/abs/2403.07918 accessed 22 March 2024; Mouton, Lucas and Guest (n 19).

[420] Lydia Morrish, 'A Face Recognition Site Crawled the Web for Dead People's Photos' Wired https://www.wired.com/story/a-face-recognition-site-crawled-the-web-for-dead-peoples-photos/ accessed 16 May 2024.

[421] Johnson and others (n 384).

[422] Darren Loucaides, 'How Governments Are Using Facial Recognition to Crack down on Protesters' [2024] *Rest of World* https://restofworld.org/2024/facial-recognition-government-protest-surveillance/ accessed 16 May 2024."

[423] Avi Asher-Schapiro and David Sherfinski, '"Scary and Chilling": AI Surveillance Takes U.S. Prisons by Storm' Reuters (16 November 2021) https://www.reuters.com/article/idUSKBN2I01GZ/ accessed 16 May 2024.

[424] Richard A Bales and Katherine VW Stone, 'The Invisible Web at Work: Artificial Intelligence and Electronic Surveillance in the Workplace' (2020) 41 Berkeley Journal of Employment and Labor Law 1.

AI models have deep censorship implications. AI models are widely used in online content moderation, enabling platforms to rapidly flag and remove inappropriate content.[425] However, content moderation AI can easily be repurposed for censorship. Many countries require use of automated content moderation tools, and several require these tools to censor political expression.[426]

AI censorship can also arise from the biases of the models themselves. Hate speech detection models have been shown to disproportionately flag language used by Black people.[427] There are several evaluations for measuring the tendency of content moderation AI to incorrectly flag speech, particularly speech from marginalised groups, as hate speech or as otherwise unacceptable.[428] While some benchmarks exist for assessing incorrect flagging of political speech as hate speech, determining this boundary is itself a political question and cannot be left entirely to evaluations.[429]

Data and content moderation labour

High quality training data is vital for reducing toxic, biased and incorrect outputs of foundation models.[430]

Companies have increasingly relied on human workers to filter out bad or toxic data, and rate and improve model outputs. These data workers are exposed to exceptionally disturbing content, including graphic images, text

[425] 'Meta's New AI System to Help Tackle Harmful Content' (*Meta*, 8 December 2021) https://about.fb.com/news/2021/12/metas-new-ai-system-tackles-harmful-content/ accessed 27 June 2024.

[426] Allie Funk, Adrian Shahbaz and Kian Vesteinsson, 'The Repressive Power of Artificial Intelligence' (Freedom House 2023) https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence accessed 16 May 2024.

[427] Maarten Sap and others, 'The Risk of Racial Bias in Hate Speech Detection' in Anna Korhonen, David Traum and Lluís Màrquez (eds), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics 2019) https://aclanthology.org/P19-1163 accessed 27 June 2024; Thomas Davidson, Debasmita Bhattacharya and Ingmar Weber, 'Racial Bias in Hate Speech and Abusive Language Detection Datasets' in Sarah T Roberts and others (eds), *Proceedings of the Third Workshop on Abusive Language Online* (Association for Computational Linguistics 2019) https://aclanthology.org/W19-3504 accessed 27 June 2024.

[428] Thomas Hartvigsen and others, 'ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection' in Smaranda Muresan, Preslav Nakov and Aline Villavicencio (eds), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Association for Computational Linguistics 2022) https://aclanthology.org/2022.acl-long.234 accessed 26 June 2024.

[429] Fabio Poletto and others, 'Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review' (2021) 55 Language Resources and Evaluation 477.

[430] Amelia Glaese and others, 'Improving Alignment of Dialogue Agents via Targeted Human Judgements' (arXiv, 28 September 2022) http://arxiv.org/abs/2209.14375 accessed 22 March 2024; Aakanksha Chowdhery and others, 'PaLM: Scaling Language Modeling with Pathways' (2023) 24 Journal of Machine Learning Research 1.

and video of torture, murder and rape.[431] They regularly face similar harms to social media content moderators such as severe emotional and social harms. This is the result of constant exposure to extremely disturbing content, frequently compounded by difficult work conditions with poor mental health support.[432] In addition, data worker jobs are often insecure and underpaid, subject to arbitrary algorithmic or remote supervision.[433]

Several evaluations for data worker conditions exist, including CrowdWorkSheets, Fairwork Principles, Criteria for Fairer Microworks and Datasheets for Datasets.[434] These evaluations give high level recommendations for better working conditions and documenting the labour and working conditions of data workers. They do not provide means for evaluating the severity of harms of different types of data work.

Labour and economic impacts

Foundation models continue long trends of disruption and marginalisation of labour by technology.[435]

By automating part or all of different jobs, these technologies reduce labour power, thereby reducing wages and power to set good working conditions. These harms are presently being experienced by artists,

[431] Casey Newton, 'The Secret Lives of Facebook Moderators in America' *The Verge* (25 February 2019) https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona accessed 22 March 2024.

[432] Karen Hao and Deepa Seetharaman, 'Cleaning Up ChatGPT Takes Heavy Toll on Human Workers' *Wall Street Journal* (24 July 2023) https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483 accessed 22 March 2024.

[433] Karen Hao and Andrea Paola Hernández, 'How the AI Industry Profits from Catastrophe' *MIT Technology Review* (20 April 2022) https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/ accessed 22 March 2024.

[434] Janine Berg and others, *Digital Labour Platforms and the Future of Work: Towards Decent Work in the Online World* (International Labour Organization 2018); Gebru and others (n 227); Mark Diaz and others, 'CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation', 2*022 ACM Conference on Fairness, Accountability, and Transparency* (2022) http://arxiv.org/abs/2206.08931 accessed 17 March 2023; 'Homepage' (Fairwork, 2024) https://fair.work/en/fw/homepage/ accessed 22 March 2024.

[435] Brian Merchant, *Blood in the Machine: The Origins of the Rebellion against Big Tech* (First edition, Little, Brown and Company 2023); Meredith Whittaker, 'Origin Stories: Plantations, Computers, and Industrial Control' [2023] Logic(s) Magazine https://logicmag.io/supa-dupa-skies/origin-stories-plantations-computers-and-industrial-control/ accessed 22 March 2024.

journalists and several other professions.[436] Content recommendation and moderation algorithms biased against marginalised people can reduce their visibility and views, impacting creator revenue.[437]

AI models can also deny economic opportunities. Biased models can provide marginalised communities with fewer or worse adverts for jobs or rentals,[438] or discriminate when ranking job applicant resumes.[439] Mortgage approval algorithms have been shown to be biased against Black applicants, approving them at lower rates than similar white applicants.[440]

There are a limited number of task-specific evaluations assessing when an AI system outperforms humans.[441] There are also several economic models and studies on job replacement by AI systems.[442] Existing studies on impacts of automation also point towards increasing inequality from increasing automation.[443] However, fully evaluating the impacts of AI systems on labour markets requires evaluations that are specific to the tasks and labour being considered. They must also consider complex social, economic and political forces shaping AI system adoption, and model improvements in AI system capabilities.

[436] David Bauder, 'Sports Illustrated Found Publishing AI Generated Stories, Photos and Authors' PBS *NewsHour* (29 November 2023) https://www.pbs.org/newshour/economy/sports-illustrated-found-publishing-ai-generated-stories-photos-and-authors accessed 22 March 2024; Pranshu Verma and Gerrit De Vynck, 'ChatGPT Took Their Jobs. Now They Walk Dogs and Fix Air Conditioners.' *Washington Post* (5 June 2023) https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs/ accessed 22 March 2024; Tom Carter, 'Workers Are Worried about AI Taking Their Jobs. Artists Say It's Already Happening.' *Business Insider* (1 October 2023) https://www.businessinsider.com/ai-taking-jobs-fears-artists-say-already-happening-2023-10 accessed 22 March 2024.

[437] Brooke Erin Duffy and Colten Meisner, 'Platform Governance at the Margins: Social Media Creators' Experiences with Algorithmic (in) Visibility' (2023) 45 Media, Culture & Society 285.

[438] Renee Shelby and others, 'Sociotechnical Harms: Scoping a Taxonomy for Harm Reduction' (arXiv, 11 October 2022) http://arxiv.org/abs/2210.05791 accessed 30 January 2023.

[439] Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women' Reuters (11 October 2018) https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/ accessed 26 June 2024.

[440] Emmanuel Martinez and Lauren Kirchner, 'The Secret Bias Hidden in Mortgage-Approval Algorithms – The Markup' [2021] *The Markup* https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms accessed 26 June 2024.

[441] Abigail See and others, 'Do Massively Pretrained Language Models Make Better Storytellers?' in Mohit Bansal and Aline Villavicencio (eds), *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (Association for Computational Linguistics 2019) https://aclanthology.org/K19-1079 accessed 26 June 2024.

[442] Eloundou and others (n 50); Xiang Hui, Oren Reshef and Luofeng Zhou, 'The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market' (31 July 2023) https://papers.ssrn.com/abstract=4527336 accessed 22 March 2024.https://papers.ssrn.com/abstract=4527336 accessed 22 March 2024.

[443] Daron Acemoglu and Pascual Restrepo, 'Tasks, Automation, and the Rise in U.S. Wage Inequality' (2022) 90 Econometrica 1973.

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

Preferred citation: Ada Lovelace Institute, *Under the radar: Examining the evaluation of foundation models* (2024)
https://www.adalovelaceinstitute.org/report/under-the-radar/

ISBN: 978-1-7395236-3-3