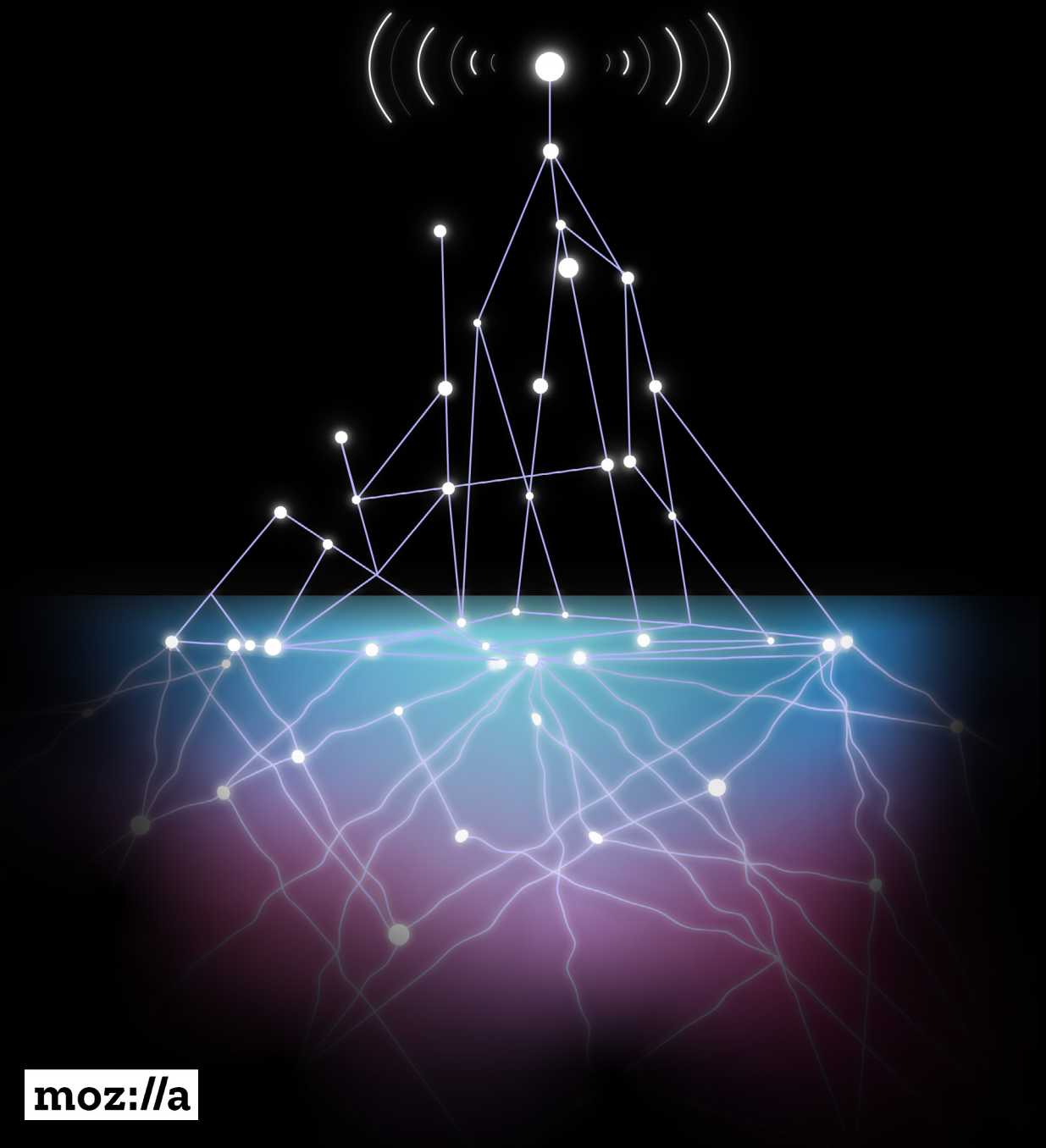
External researcher access to closed foundation models

State of the field and options for improvement

ESME HARRINGTON & DR. MATHIAS VERMEULEN (AWO)

SUPPORTED BY THE MOZILLA FOUNDATION



moz://a

External researcher access to closed foundation models:
State of the field and options for improvement

© 2024
by Esme Harrington & Mathias Vermeulen
is licensed under CC BY-SA 4.0. To view a copy of this licence,
visit http://creativecommons.org/licenses/by-sa/4.0

External researcher access to closed foundation models

State of the field and options for improvement

ESME HARRINGTON & DR. MATHIAS VERMEULEN (AWO)

21 AUGUST 2024[1]

1 The majority of research, including interviews, was conducted between March and April 2024. We updated the public policy analysis and industry mapping in August 2024.

Table of Contents

Executive Summary

As foundation models become increasingly embedded in a wide array of downstream products and services, understanding their risks and vulnerabilities is more critical than ever to prevent negative impacts. External scrutiny can play a crucial role not only in forming a comprehensive understanding of these risks and vulnerabilities but also in ensuring that users, regulators, and the general public can trust that a foundation model has been rigorously tested.

This raises questions concerning the minimum conditions for public scrutiny and public-interest research for those who choose to keep their model gated behind APIs or proprietary interfaces, including most dominant firms in the industry. Current policy initiatives in the EU, UK, and US have addressed this question only to a limited extent. The EU's AI Act introduces specific legal obligations for developers of foundation models, including red teaming and risk assessments, but falls short of spelling out minimum conditions. The UK and the US have included proposals around external researcher access into various non-binding policy frameworks, without mandating any form of external access.

Based on desk research and interviews with researchers, this report examines the status quo of external researcher access amongst several of the leading closed foundation model and generative AI developers, (a) OpenAI (GPT4), (b) Google (Gemini), (c) Cohere (Command), (d) Anthropic (Claude), (e) Midjourney, and (f) Inflection (Pi).

We find that these voluntary initiatives are taking a varied and often insufficient approach to facilitate public-interest research.

- Overall, developers are gatekeepers of access programs, which enables them to prioritise research that aligns with internal priorities or is less commercially threatening. Compounding concerns, selection processes may not be sufficiently resourced or transparent enough to promote trust.
- Some developers involve external researchers in pre-release evaluations and red teaming, predominantly where it aligns with internal risk priorities.
- Robust external research on foundation models can be expensive, creating an inequitable barrier to entry for researchers. To address this, some developers offer subsidised API access for researchers.
- Developers use terms of service enforcement processes that may suspend or terminate the accounts of researchers conducting evaluations or red teaming, and few developers operate sufficient appeals processes.

- Current safe harbours in developers' vulnerability reporting and disclosure programs may not be sufficient to prevent legal risks from having a chilling effect on research.
- Most developers do not offer sufficiently transparent model versioning or model stability, including usage of additional components and filters, which hinders reproducible research.
- Model evaluations require access to conduct sampling and fine-tuning, which are available via many developers' APIs. Yet researchers often don't have access to base models, model families, and components such as filters and content moderation systems. Other research may require deeper levels of access to models, internal data, and documentation, which are currently unavailable.
- Leading developers do not provide access to training data due to legal and reputational risks as well as for competitive reasons.
- Developers do not share any information about usage trends.
- Developers are not sufficiently transparent in supporting documentation, particularly in relation to content moderation, environmental, and labour practices.

As a result, we propose the following range of activities to improve this status quo, spanning from practical changes to current voluntary programmes to ambitious long-term goals that will likely require policy intervention:

Recommendations for industry:

1. Provide researchers with subsidised API access.
2. Provide transparent and effective content moderation and appeals processes, with fast-track appeals for researchers.
3. Develop comprehensive safe harbour and responsible disclosure policies.
4. Provide clear and accurate model versioning and stability.
5. Provide researchers with sufficient levels of model and data access via a structured access program.
6. Develop crowdsourced data collection approaches for foundation model research, including to collect usage information.
7. Publish transparency reports.

Recommendations for policy makers:

8. Create an independently mediated structured researcher access program.

1.
Introduction

Section 1: Introduction

The recent buzz around foundation models has, among many other things, raised the question of how open and accessible such models, especially the most capable models, should be — leading to a hotly contested debate around "open" vs. "closed" AI systems (and all the gradient of release in between).[2] This also raises questions concerning the minimum conditions for public scrutiny and public-interest research for those who choose to keep their model gated behind APIs or proprietary interfaces, including most dominant firms in the industry.

With increased integration into downstream products and services, gaining a better understanding of the risks and vulnerabilities of such models will only become more important in order to mitigate potential negative impacts down the line. At the same time, enhanced researcher access to foundation models can also help evaluate model safety and mitigation features used to make models safer.

Recognising this, the EU, UK, and US have introduced legislation and non-binding policy frameworks that recognise the importance of AI safety and security research, whilst falling short of mandating external researcher access. Several jurisdictions have also created public AI Safety Institutes to facilitate or themselves conduct external evaluations of leading foundation models. However, ensuring robust research and oversight of foundation models requires a diversity of expertise, objectives, and actors. Research conducted by a range of external researchers is an essential complement to public bodies' evaluations and developers' internal due diligence, adding an additional layer of accountability.

Against this background, this paper explores the following questions: (1) What is the status quo on external researcher access among the following developers: (a) Open AI (GPT4), (b) Google (Gemini), (c) Cohere (Command), (d) Anthropic (Claude), (e) Midjourney, and (f) Inflection (Pi),[3] (2) What are researchers concerned about within this status quo? (3) Which activities could better enable external research? This report is based on desk research and interviews with external researchers with expertise on foundation models and/or technology researcher access programs.

2 Solaiman, The Gradient of Generative AI Release: Methods and Considerations, 05 February 2023, https://arxiv.org/pdf/2302.04844.pdf

3 We chose to focus on these developers because they are some of the leading foundation model developers which offer some form of company-mediated researcher access initiatives or harm reporting tools as identified by Longpre et al., A Safe Harbor for AI Evaluation and Red Teaming.

2.
Definitions

Section 2: Definitions

Foundation models, in the simplest meaning of the term, are AI models capable of a range of general tasks such as generating text, images, and audio.[4] They often act as a base for a range of downstream applications that are built on top of or fine-tuned from the foundation model. These applications may be deployed by the same developer that developed the foundation model (e.g. OpenAI provides the chatbot ChatGPT) or by a different actor.

External researchers in this paper refers to any researcher not working in-house for a developer either on a not-for-profit basis or for a recognised public interest mission. Under relevant EU law, for example in the copyright directive, external researchers are defined by affiliation to a research organisation, namely a university, research institute, or other entity whose primary goal is to conduct scientific research or to carry out education activities involving the conduct of scientific research, either on a not-for-profit basis or for a recognised public interest mission.[5] This definition may not capture the full spectrum of external researchers, excluding those affiliated with an organisation with additional objectives, such as advocacy work (which is more likely to be the case for researchers focused on social impact work).

External research on foundation models can take many forms including auditing, evaluations and red teaming, and broader social impacts research.[6] These terms are currently being concretised in a variety of fora, including the EU-US Trade and Technology Council.[7] AI auditing encapsulates a range of processes that involve "independent evaluations of the performance, fairness or safety of deployed AI systems."[8] In the context of foundation models, evaluations refer to the assessments of a model's safety and security relevant properties.[9] Red teaming refers to "structured testing effort to find flaws and vulnerabilities in an AI

4 Jones, Explainer: What is a foundation model?, Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

5 Article 2(1) Copyright Directive (EU) 2019/790, https://eur-lex.europa.eu/eli/dir/2019/790/oj; the Digital Services Act also uses this definition.

6 Anderljung, Thornton Smith, O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 November 2023, https://arxiv.org/abs/2311.14711

7 EU-US Terminology and Taxonomy for Artificial Intelligence, 31 May 2023, https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence

8 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861, p. 1.

9 Marsh, Introducing the AI Safety Institute, AI Safety Institute, DSIT, November 2023, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

system, often in a controlled environment and in collaboration with developers of AI."[10] This may occur by a red team, a group that are "authorised and organised to emulate a potential adversary's attack or exploitation capabilities"[11] or by external researchers that are not pre-authorised and should instead disclose vulnerabilities through reporting channels.

Research domains for foundation models are varied and range from fairness, safety, security, and privacy,[12] to disinformation.[13] External research primarily takes place once models have been released onto the market. However, it can occur earlier in the product development lifecycle with some developers seeking domain experts to conduct evaluations and red teaming prior to release. Depending on the product development stage and access point, external research may be authorised and facilitated by developers (i.e. 'second party') or be fully independent from the developer (i.e. 'third party').[14]

External researchers can act both as pathfinders and quasi-auditors, spotting new and emerging risks and verifying developer claims.[15] In particular, they can identify blind spots, biases, or vulnerabilities in the model that internal teams might overlook due to their familiarity with the system's design and assumptions.[16] External researchers may also possess specialised knowledge or expertise in areas that the internal team might not. To ensure a diversity of expertise and prevent bottlenecks, external research cannot be allocated to any one external organisation (such as an AI Safety Institute). Overall, involving a range of external researchers demonstrates a commitment to transparency and accountability, reassuring stakeholders, including users, regulators, and the public, that the foundation model has been rigorously tested and vetted by independent experts.[17]

10 The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 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/

11 EU-US Terminology and Taxonomy for Artificial Intelligence, 31 May 2023, https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence, p. 12.

12 Das, Amini, Wu, Security and Privacy Challenges of Large Language Models: A Survey, 30 January 2024, https://arxiv.org/abs/2402.00888

13 Center for Countering Digital Hate, Fake Image Factories, 06 March 2024, https://counterhate.com/research/fake-image-factories/

14 Constanza-Chock, Harvey, Raji, Czernuszenko, Boulamwini, Who audits the auditors? Recommendations from a field scan of the algorithmic auditing ecosystem, 04 October 2023, https://arxiv.org/abs/2310.02521

15 Vermeulen, Researcher Access to Platform Data: European Developments, 09 September 2022, https://tsjournal.org/index.php/jots/article/view/84; Anderljung, Thornton Smith, Joe O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 November 2023, https://arxiv.org/abs/2311.14711

16 Inflection, Our policy on frontier safety, 30 October 2023, https://inflection.ai/frontier-safety

17 National Science Foundation, Democratizing the future of AI R&D: NSF to launch National AI Research Resource pilot, 24 January 2024, https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai

3.

Emerging policy in the US, UK, and EU

Section 3: Emerging policy in the US, UK, and EU

3.1 EU Regulation

The European Union's AI Act introduced specific obligations for general-purpose AI systems (GPAI), often also referred to as foundation models.[18] Providers of GPAI have several due diligence obligations, including to conduct red teaming testing and systemic risk assessments if the GPAI qualifies as a GPAI with systemic risk.[19] In addition, providers will need to create and maintain technical documentation that includes information on the training and testing process of the model.[20] This should include a clear listing and description of the datasets and the data processing techniques used throughout model training.[21] The European Commission will adopt delegated acts to supplement guidance and the benchmarks and indicators necessary for assessing the risk of general purpose AI.[22] The EU AI Office has the exclusive competence to oversee and enforce these obligations.

Foundation models that are deployed by very large online platforms as a product feature could also be covered by Article 40 of the Digital Service Act. This means that foundation model-based products deployed by Microsoft, Google, and Meta could fall in-scope.[23] For example, Inflection's Pi is available via Instagram, Facebook Messenger, and WhatsApp. Article 40(4) of the DSA creates a structured researcher access infrastructure that enables researchers to request a wide range of platform data to conduct research on systemic risks. AlgorithmWatch and AI Forensics sent a request for data under Article 40(4) to Microsoft, including data related to public usage of Microsoft's Co-Pilot which is built on GPT-4.[24]

18 Jones, Explainer: What is a foundation model?, Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

19 AI Act Article 55(1)(a) and (b).

20 AI Act Article 52c.1(a)

21 Sufficiently detailed? A proposal for implementing the the AI Act's training data transparency requirements for GPAI, Open Future and Mozilla Foundation, June 2024, https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf.

22 AI Act Article 52a.3.

23 Lemoine, Laureline, Vermeulen, Mathias, From Chat GPT to Google's Gemini – when would generative AI products fall within the scope of the Digital Services Act. London School of Economics, 12 February 2024, https://blogs.lse.ac.uk/medialse/2024/02/12/from-chatgpt-to-googles-gemini-when-would-generative-ai-products-fall-within-the-scope-of-the-digital-services-act/

24 AlgorithmWatch, Got Complaints? Want Data? Digital Service Coordinators will have your back – or will they? 14 February 2024, https://algorithmwatch.org/en/dsa-day-and-platform-risks/

3.2 Policy Initiatives and Voluntary Commitments in the UK and US

In October 2023, the UK hosted the first AI Safety Summit, convening governments, industry, and other stakeholders to consider the risks of advanced foundation models. The countries in attendance signed the Bletchley Declaration which committed to developing scientific research in relation to independent safety testing.[25] At a session dedicated to safety testing, governments also agreed to work with other appropriate external organisations to conduct this testing.[26]

Just before the AI Safety Summit, the US administration issued an Executive Order on the safe, secure, and trustworthy development and use of AI,[27] which emphasised the importance of red teaming. The White House has used the Defense Production Act to compel developers to share vital information, including safety test results, with the Department of Commerce.[28] The Executive Order also requested the National Institute of Standards and Technology (NIST) to develop red teaming guidance.

In July 2023, the White House secured voluntary commitments from Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI — additional companies have joined later on — concerning the safe development and deployment of their foundation models. In particular, these developers committed to 1) pre-release "internal and external security testing" of models and 2) to "facilitat[e] third-party discovery and reporting of vulnerabilities."[29] In particular, they committed to conduct internal and external security testing of systems prior to release, partly by external experts, on significant sources of risks such as biosecurity,

25 DSIT, FCDO, 10 Downing Street, The Bletchley Declaration by Countries Attending the AI Safety Summit,01 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

26 DSIT, FCDO, 10 Downing Street, Safety Testing: Chair's Statement of Session Outcomes, 02 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-safety-testing-2-november/safety-testing-chairs-statement-of-session-outcomes-2-november-2023

27 The White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 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/

28 The White House, Fact Sheet: Biden-Harris Administration Announces Key AI Actions Following President Biden's Landmark Executive Order, 29 Janaury 2024, https://www.whitehouse.gov/briefing-room/statements-releases/2024/01/29/fact-sheet-biden-harris-administration-announces-key-ai-actions-following-president-bidens-landmark-executive-order/

29 The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manages the Risks Posed by AI, 21 July 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

cybersecurity, and broader societal effects.[30] Following this, several developers have published voluntary statements committing to further work with external labs and evaluation organisations prior to model release.[31]

The United States' National Telecommunications and Information Administration's (NTIA)[32] report on AI Accountability Policy also discussed the need for fostering an independent audit ecosystem, noting that commenters "urged the government to facilitate appropriate external access to AI systems." [33] The UN interim report on Governing AI for Humanity suggests an international mechanism is needed to facilitate access to compute and other resources for external researchers to conduct research and evaluations.[34] The final UN report will be an input to the Pact for the Future and the Global Digital Compact which are being negotiated between member states.[35]

3.3 AI Safety Institutes

Several countries have created quasi-regulatory bodies, called AI Safety Institutes, to work on the safety and security of foundation models and other AI systems. Each has a different remit, from conducting in-house evaluations to facilitating multi-stakeholder development of best practices.

During the UK AI Safety Summit, the national signatories of the Bletchley Declaration also committed to developing public sector capability in relation to independent safety testing.[36]

30 The White House, FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, 21 July 2023, https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

31 Anthropic, Frontier Threats Red Teaming for AI Safety, 26 July 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety#entry:146918@1:url; Our policy on frontier safety, Inflection, 30 October 2023, https://inflection.ai/frontier-safety

32 The NTIA is housed in the United States Department of Commerce.

33 Goodman, Artificial Intelligence: Accountability Policy Report, National Telecommunications and Information Administration, March 2024, https://www.ntia.gov/sites/default/files/publications/ntia_ai_report_final-3-27-24.pdf, p.35.

34 UN and AI Advisory Body, Interim Report: Governing AI for Humanity, December 2023, https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf, p. 20.

35 Harrington, Interview with Filippo Pierozzi in Algorithm Governance Roundup #13, AWO, https://eocampaign1.com/web-version?p=b54cf9a4-f1a0-11ee-b59d-cdcd2559006b&pt=campaign&t=1712165248&s=1857b9d891aabf35f0a-9dac3097e59a4c38e3a2b1f6ec63b8be580c264595f03

36 DSIT, FCDO, 10 Downing Street, The Bletchley Declaration by Countries Attending the AI Safety Summit, 01 November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023

Since then, the United Kingdom,[37] the US,[38] Canada,[39] and Japan[40] have set up AI Safety Institutes (AISIs). Following the passage of the EU's AI Act, the EU's AI Office has been set up to implement the legislation and has exclusive competence for GPAI/foundation models. The EU AI Office is developing close relationships with the AISIs to share best practices, including through the EU-US Trade and Technology Council.[41]

In April 2024, the UK and US AISIs announced a partnership to research and evaluate foundation model safety.[42] The two bodies will collaborate on the development of tests for foundation model safety and perform at least one joint testing exercise on a publicly accessible foundation model. At the same time, the EU-US Trade and Technology Council announced a joint roadmap to develop common metrics and benchmarks to assess the trustworthiness and risk management of AI systems.[43] In May 2024, the UK further announced it was opening an office in San Francisco,[44] as well as partnering with the Canadian AISI.[45]

3.3.1 UK AI Institute

The UK AISI is an in-house government research centre dedicated to safety research and conducting external evaluations of advanced AI including foundation models.[46] In September 2023, the UK's AISI secured voluntary agreements with OpenAI, Google, Microsoft, Anthropic, and Meta to conduct safety testing on their advanced foundation models prior to de-

37 Marsh, Introducing the AI Safety Institute, AI Safety Institute, DSIT, November 2023, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

38 US Artificial Intelligence Safety Institute, National Institute of Standards and Technology, 08 February 2024, https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

39 Department of Finance Capital, Remarks by the Deputy Prime Minister on securing Canada's AI advantage, 07 April 2024, https://www.canada.ca/en/department-finance/news/2024/04/remarks-by-the-deputy-prime-minister-on-securing-canadas-ai-advantage.html

40 Japan AI Safety Institute, About Us, https://aisi.go.jp/home/aboutus/

41 European Commission, EU-US Trade and Technology Council (2021-2024), https://digital-strategy.ec.europa.eu/en/factpages/eu-us-trade-and-technology-council-2021-2024

42 UK and United States announce partnership on science of AI safety, DSIT, AI Safety Institute, 02 April 2024, https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety ; Manancourt, Volpicelli, Chatterjee, 'Rishi Sunak promised to make AI safe. Big Tech's not playing ball., Politico Pro, 26 April 2024, https://pro.politico.eu/news/178741

43 https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management

44 AI Safety Institute, DSIT, Government's trailblazing Institute for AI Safety to open doors in San Francisco, 20 May 2024, https://www.gov.uk/government/news/governments-trailblazing-institute-for-ai-safety-to-open-doors-in-san-francisco.

45 AI Safety Institute, DSIT, UK-Canada science of AI safety partnership, 20 May 2024, https://www.gov.uk/government/publications/uk-canada-science-of-ai-safety-partnership/uk-canada-science-of-ai-safety-partnership/.

46 It was built on and replaced the UK government's Frontier Model Taskforce.

ployment in relation to misuse, societal impacts, autonomous systems, and safeguards.[47] To support this work, the UK AISI agreed several partnerships with private sector AI evaluation organisations.[48] Despite this, the AISI has struggled to obtain pre-release access to conduct safety testing, with OpenAI, Anthropic, and Meta releasing new foundation models without granting access.[49] Recently, Google and Anthropic have allowed the AISI to conduct some tests on their most advanced models prior to full release (Gemini and Clause 3.5 Sonnet respectively).[50]

Developers are reportedly hesitant to grant the UK AISI pre-deployment access on the basis that it could set a precedent in other countries.[51] Developers have also put their reluctance down to a lack of transparency from the UK AISI, requesting further details about how the AISI are conducting tests, how long they will take, and how the feedback process will work.[52] The recent partnership with the US AISI aims to encourage developers to provide pre-release access.[53] The UK government has indicated that they may introduce mandatory safety and information sharing requirements on leading AI developers and place the AISI on a statutory footing.[54]

3.3.2 US AI Safety Institute

The US AISI is housed under NIST, and aims to bring together AI developers and users, academics, government, industry researchers, and civil society to collaboratively develop guidance, methods, and best practices.[55] In February 2024, the US AI Safety Institute Consortium (AISIC) was created to develop guidelines for red teaming, capability evaluations,

47 Milmo and Stacey, Tech firms to allow vetting of AI tools, as Musk warns all human jobs threatened, 03 November 2023, https://www.theguardian.com/technology/2023/nov/02/top-tech-firms-to-let-governments-vet-ai-tools-sunak-says-at-safety-summit ; AI Safety Institute:third progress report, DSIT, 05 February 2024, https://www.gov.uk/government/publications/uk-ai-safety-institute-third-progress-report/ai-safety-institute-third-progress-report

48 Frontier AI Taskforce brings in leading technical organisations to research risks, DSIT, 18 October 2023, https://www.gov.uk/government/news/frontier-ai-taskforcebrings-in-leading-technical-organisations-to-research-risks

49 Manancourt, Volpicelli, Chatterjee, 'Rishi Sunak promised to make AI safe. Big Tech's not playing ball., Politico Pro, 26 April 2024, https://pro.politico.eu/news/178741

50 Ibid; Expanding access to Claude for government, Anthropic, 26 June 2024, https://www.anthropic.com/news/expanding-access-to-claude-for-government.

51 Ibid.

52 Criddle, Gross and Murgia, World's biggest AI tech companies push UK over safety tests, 07 February 2024, https://www.ft.com/content/105ef217-9cb2-4bd2-b843-823f79256a0e

53 Manancourt, Volpicelli, Chatterjee, 'Rishi Sunak promised to make AI safe. Big Tech's not playing ball., Politico Pro, 26 April 2024, https://pro.politico.eu/news/178741

54 Bambridge, Politico Pro, UK to consult on copyright regime as it prepares AI legislation, 30 July 2024, https://pro.politico.eu/news/183938

55 US Artificial Intelligence Safety Institute, National Institute of Standards and Technology, 08 February 2024, https://www.nist.gov/artificial-intelligence/artificial-intelligence-safety-institute

risk management, safety, and security.[56] The AISIC involves industry, state and local government, civil society, and academia.

The US AISI has not yet conducted external evaluations. Instead, the Executive Order launched a pilot National Artificial Intelligence Research Resource (NAIRR) within the National Science Foundation which disperses industry-donated computational (and other) resources to external researchers who conduct evaluations and other research.[57] However, in August 2024, the AISI approached several companies, including Anthropic and Meta,[58] and secured agreement from OpenAI to conduct pre-release safety testing.[59] This indicates a shift from developing guidance towards in-house safety testing, similar to the UK AISI. Given that many leading developers are headquartered in the U.S., they appear to have stronger relationships with the US AISI, as opposed to the UK AISI.

56 Biden-Harris Administration Announces First-Ever Consortium Dedicated to AI Safety, National Institute of Standards and Technology, 08 February 2024, https://www.nist.gov/news-events/news/2024/02/biden-harris-administration-announces-first-ever-consortium-dedicated-ai

57 National Artificial Intelligence Research Resource, https://new.nsf.gov/focus-areas/artificial-intelligence/nairr

58 Chatterjee, Politico Pro, US AI Safety Institute Trying to Secure More Testing Agreements, 13 August 2024, https://pro.politico.eu/news/184442

59 Wiggers, OpenAI pledges to give U.S. AI Safety Institute early access to its next model, 31 July 2024, https://techcrunch.com/2024/07/31/openai-pledges-to-give-u-s-ai-safety-institute-early-access-to-its-next-model/

4.

Status quo of researcher access

Section 4 : Status quo of researcher access

Except for the EU's AI Act, the majority of these emerging policy initiatives highlighted in section 3 are voluntary, which leaves developers free to engage on their own terms.

Longpre et al. have analysed several of the leading foundation model developers' initiatives, policies, and processes to assess how they impact external research.[60] Building on this work, this section explores these initiatives and policies, researchers' concerns, and assesses how they measure up to OpenAI, Google, Anthropic, Inflection, Midjourney, and Cohere's voluntary commitment to facilitate third-party discovery and reporting of model vulnerabilities.[61] Annexes 2, 3, and 4 provide granular detail on developers' policies.

AI Company	AI System	Public API / Open	Deep Access	Researcher Access	Bug Bounty	Safe Harbor	Enforcement Process	Enforcement Justification	Enforcement Appeal
OpenAI	GPT-4	●	◑	●	●	◑†	●	○	◑
Google	Gemini	●	○	○	●	○	○	◑	○
Anthropic	Claude 2	○	○	◑	○	◑‡	●	○	○
Inflection	Inflection-1	○	○	○	○	○	○	◑	◑
Meta	Llama 2	●	●	●	●	◑‡	○	○	○
Midjourney	Midjourney v6	○	○	○	○	○	○	○	◑
Cohere	Command	●	○	●	○	◑	○	○	○

Table 3. **A summary of the policies, access, and enforcement for major AI systems, suggesting a challenging environment for independent AI research.** We catalog if each system has a public API, deeper access than final outputs (e.g. top-5 logits for OpenAI), researcher access programs, security research bug bounties, any legal safe harbors, and whether they disclose their account enforcement process, disclose justification on enforcement actions, and have an enforcement appeals process. ● indicates the company satisfies this criteria; ○ indicates it does not, and ◑ indicates partial satisfaction. ‡ Indicates security-only research safe harbors, "solely at [their] discretion". † Indicates a safe harbor for security and "academic research related to model safety". The latter was added by OpenAI in response to reading an early draft of this proposal, though some ambiguity remains as to the scope of protected activities. Full details are provided in Table A1.

Image taken from: A Safe Harbor for AI Evaluation and Red Teaming, Longpre, Kapoor, Klyman et al, 2024.

60 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893

61 We decided not to include Meta's Llama 2 because the model, including weights, are freely available to download locally. As a result, researcher access is not gated behind company infrastructure putting it out of scope of this paper. To note, Llama 2 does not qualify as open source under the Open-Source Initiative's definition.

4.1 Pre-release access

> Some developers involve external researchers in pre-release evaluations and red teaming processes, predominantly where it aligns with internal risk priorities.

Annex 1 compares pre-release evaluation and red teaming processes which have involved external researchers. These have been documented to occur at OpenAI, Anthropic, and Inflection. Both Anthropic and Inflection involved experts from domains that aligned with internal risk priorities, particularly biosecurity. In addition, Anthropic involves subject matter experts for trust and safety, national security, and multilingual and multicultural red teaming throughout the model lifecycle. In December 2023, OpenAI formalised its approach and publicly recruited for an ongoing Red Teaming Network.[62] These second-party evaluations are subject to strong expectations of confidentiality because developers are concerned about model capabilities being leaked. It is unclear if other developers are meeting this commitment due to a lack of publicly available information.

4.2 Cost of access and subsidised API initiatives

> Robust external research on foundation models can be expensive, creating an inequitable barrier to entry for researchers.

Much of the external research on closed foundation models is conducted by third party researchers via commercial APIs and at the public application layer.[63] APIs are a form of transparency infrastructure, described by Ojewale et al. as an interface hosted by model operators that allows controlled access to proprietary information about a model.[64] Researchers evaluating foundation models will often need to conduct sampling[65] many times in an automated and systematic manner. As a result, they may need to use commercial APIs with higher rate-limits which can become expensive.

62 OpenAI Red Teaming Network, OpenAI, https://openai.com/blog/red-teaming-network

63 Many of the largest foundation model developers publicly deploy downstream applications such as chatbots and playgrounds. Jones, Explainer: What is a foundation model?, Ada Lovelace Institute, 17 July 2023, https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

64 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861

65 Sampling involves submitting a prompt and observing the models output per Bucknall and Trager, page 6.

Across these developers, OpenAI, Google, and Cohere offer public downstream applications which are available for free. These are subject to rate limits, limiting the number of prompt and output tokens per user. Commercial APIs are available at tiered rate limits for all of these developers except Midjourney, which is only available as a paid for proprietary application.

> Some developers offer subsidised API access to external researchers upon direct application or through the US's National AI Research Resource (NAIRR) Pilot. This improves access for resource-constrained researchers and encourages third party discovery of vulnerabilities.

Annex 2 compares researcher access programs across all the in-scope developers. OpenAI, Anthropic, Cohere, and Inflection offer a program for researchers to apply for subsidised access to model APIs. To apply, external researchers must submit an application explaining the research use case and their individual profile. The selection processes are somewhat unclear and the timeline for review is around every four weeks at Anthropic and four to six weeks at OpenAI.

In addition, Anthropic, Google, Cohere, and OpenAI provided testing models to a large public generative AI red teaming event at DEFCON 2023.[66] Of note, the testing APIs provided were specifically designed to not lead to bans for inappropriate prompts and had high rate-limits to handle the levels of sampling required.[67] To safeguard against abuses, the challenge had a vulnerability disclosure process and was conducted on secure laptops.

OpenAI, Microsoft (GPT-4), and Anthropic also all donated API access or compute resources to the NAIRR Pilot.[68] This publicly funded program allocates subsidised API access and computational resources to US-based researchers who apply via the National Science Foundation, detailing their research use case and researcher profile. OpenAI donated $1.25 million in model API credits for research on AI safety, evaluations, and societal impacts. Microsoft donated $20 million in compute credits and access to models including GPT-4 via Azure OpenAI. Anthropic donated API credits but they are not for model safety research.

Policy option 1:
Provide researchers with subsidised API access.

66 Cattell, Chowdhury, Carson, AI Village at DEF CONN announces largest-ever public Generative AI Red Team, AI Village, 03 May 2023, https://aivillage.org/generative%20red%20team/generative-red-team/

67 Cattell, Generative Red Team Recap, AI Village, 12 October 2023, https://aivillage.org/defcon%2031/generative-recap/.

68 https://new.nsf.gov/focus-areas/artificial-intelligence/nairr

4.3 Enforcement processes and appeals

> Developers use terms of service enforcement processes that may suspend or terminate the accounts of researchers conducting evaluations or red teaming, and few developers operate sufficient appeals processes. This hinders third party discovery of vulnerabilities.

A significant amount of foundation model research aims to train and fine-tune models to be as safe as possible. However, it is increasingly recognised that it may not be possible to make a model completely safe.[69] As a result, most of these developers have introduced additional enforcement systems at the API and application layer to monitor and moderate user behaviour. For example, OpenAI and Microsoft recently published a blog about their detection of a hacking group by monitoring user queries.[70]

All of these developers' policies prohibit use that breaches legislation, harms others, or generates harmful content, including violent or sexual content.[71] Safety evaluations and red teaming often purposefully attempt to produce outputs that breach these policies. From the developer's perspective, malicious actors and safety researchers attempting to exploit vulnerabilities look the same, which means researchers are at risk of account suspension or termination.[72]

Annex 3 maps the enforcement processes that developers have introduced to enforce usage policies, including whether justifications and appeals processes are provided. All of these developers except Cohere explain that they use some enforcement processes to monitor API or application user behaviour. OpenAI and Inflection provide a detailed explanation of the content moderation systems and human reviewers involved in their enforcement processes. OpenAI also provides appeals processes for users that have had accounts suspended or deleted. Overall, enforcement rates seem to be quite low, but Longpre et al. found that OpenAI, Anthropic, Inflection, and Midjourney have suspended user accounts

69 Narayanan and Kappor, AI Safety is not a model property, https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property.

70 Microsoft Threat Intelligence, Staying ahead of threat actors in the age of AI, 14 February 2024, https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

71 Developer usage policies; https://openai.com/policies/usage-policies; https://www.anthropic.com/legal/aup; https://ai.google.dev/docs/safety_setting_gemini; https://docs.cohere.com/docs/usage-guidelines; https://docs.midjourney.com/docs/terms-of-service

72 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893; Massachusetts Institute for Technology, Safe Harbour for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/.

conducting public interest research.[73] For example, researchers evaluating the propensity of Midjourney v6 to generate potentially copyright-infringing material had several commercial accounts terminated without refund.[74]

> *Policy option 2:*
> *Provide transparent and effective content moderation and appeals processes, with fast-track appeals for researchers.*

4.4 Vulnerability reporting and legal safe harbours

> Most developers have created harm discovery tools to support and incentivise the researchers to disclose safety and security vulnerabilities in line with voluntary commitments, such as bug bounties and vulnerability reporting programs. However, current safe harbours may not be sufficient to prevent legal risks from having a chilling effect on research.

Annex 4 maps the vulnerability reporting processes and legal safe harbours at the in-scope developers. OpenAI, Anthropic, Google, and Cohere all offer a bug bounty program for security vulnerability reporting. However, only OpenAI and Anthropic provide an explicit safe harbour for good faith security research. It is unclear whether any of the developers' safety reporting processes have a specific responsible disclosure policy or safe harbour.

Anti-hacking legislation in the US and the EU introduces legal risks for good-faith safety and security red teamers.[75] In the EU, the Directive on Attacks against Information Systems requires Member States to criminalise cyber security attacks.[76] The Directive acknowledges that external security red teaming can be conducted in the public interest but merely suggests Member States "provide possibilities for the legal detection and reporting of security gaps."[77] The recent Cyber Resilience Act has been the focus of efforts to create such a legal

73 Ibid, page 5.

74 Vincent, AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, 16 January 2023, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

75 Massachusetts Institute for Technology, Safe Harbour for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/.

76 Directive 2013/40/EU on attacks against information systems, 12 August 2013, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013L0040.

77 Ibid, recital 12.

safe harbour.[78] A comprehensive safe harbour was ultimately not included, but recitals in the final text encourage Member States to adopt "guidelines as regards the non-prosecution of information security researchers and an exemption from civil liability for their actions."[79]

The US's Computer Fraud and Abuse Act (CFAA) makes it a crime for actors to intentionally access a computer without authorisation or exceed authorised access by using the access to obtain unauthorised information.[80] In addition, the CFAA allows companies to bring civil action for loss against actors who violate its provisions.[81] In the social media context, platforms have banned essential data collection methods (such as scraping and sock puppets[82]) in their terms of service.[83] As a result, platforms have threatened independent researchers under the CFAA for using these unauthorised data collection methods, particularly web scraping.[84] Recently, the Department of Justice has published a policy stating it will not bring CFAA prosecutions against security researchers conducting solely good faith research.[85]

Given the broad nature of the CFAA provision, Longpre et al. argue that external foundation model red teamers may be at risk of both criminal and civil litigation under the legislation, and this creates a chilling effect on research.[86] As a result of this perceived threat, several developers have provided explicit legal safe harbours in their security vulnerability disclosure programs (see Annex 4).

Case law has narrowed the application of the CFAA to public interest technology research. In *Sandvig v Barr,* the Court ruled that research investigating whether online algorithms

78 The Cyber Resilience Act, How to make Europe more digitally resilient?, EDRi, https://edri.org/wp-content/uploads/2023/05/Cyber-Resilience-Act-draft-position-EDRi_final.pdf

79 Recital (35i) https://data.consilium.europa.eu/doc/document/ST-17000-2023-INIT/EN/pdf

80 US Code Section 1030 - Fraud and related activity in connection with computers, https://www.law.cornell.edu/uscode/text/18/1030#a_2

81 Loss is narrowly defined to "technological harm" (i.e a corrupted files) per *Van Buren v United States*.

82 Scraping is the process of automatically collecting information from web pages and sock puppets refer to a fictitious online identity created to conduct research without revealing the researcher's true identity, see Technical methods for regulatory inspection of algorithmic systems, Ada Lovelace Institute, https://www.adalovelaceinstitute.org/report/technical-methods-regulatory-inspection/.

83 Abdo, Krishnan, Krent, Falcón and Woods, A Safe Harbor for Platform Research, Knight First Amendment Institute, 19 January 2023, https://knightcolumbia.org/content/a-safe-harbor-for-platform-research; Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861 page 10.

84 Taylor Hatmaker, Facebook cuts off NYU researcher access, prompting rebuke from lawmakers, 04 August 2021, https://techcrunch.com/2021/08/04/facebook-ad-observatory-nyu-researchers/

85 Department of Justice Annoucnes New Policy for Charging Cases under the Computer Fraud and Abuse Act, Office of Public Affairs, 19 May 2022, https://www.justice.gov/opa/pr/department-justice-announces-new-policy-charging-cases-under-computer-fraud-and-abuse-act

86 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page.7.

result in discrimination using sock puppet accounts did not violate the CFAA.[87] In *Van Buren,* the Supreme Court held that the CFAA's provision on "exceeding authorized access" does not encompass "violations of circumstance-based access restrictions."[88] In other words, if someone has legitimate access to a database, they do not violate the CFAA by using that access for an improper purpose. Instead, the Court adopted a "gates-up-or-down" approach, hinging liability on whether an actor is entitled to access the information or not.[89] The decision narrowed the interpretation to prevent criminalising a broad range of common online activities that may violate terms of service agreements but do not constitute hacking. The Court in *Van Buren* did not rule whether an actor has to overcome a technological restriction to exceed authorised access.

It is unclear how the CFAA and EU anti-hacking law applies to safety researchers that conduct red teaming to force a model to generate violative content. If the violative content is produced using basic sampling via an authorised API or application account, it is arguable this is merely a breach of the terms of services rather than overcoming a technological 'gate' or restriction. However, this legal ambiguity does pose a chilling effect on research.

US copyright legislation, through the Digital Millennium Copyright Act (DMCA), allows for civil lawsuits if an actor circumvents technological protection measures that control access to works. As Longpre et al. note, this has been relied on by OpenAI in an attempt to dismiss a lawsuit brought by the New York Times, and researchers have submitted a petition for exemption to investigate bias in generative AI systems.[90] This is likely more of a risk for multi-modal foundation models that can generate images such as Midjourney. Indeed, Midjourney specifically prevents use that generates copyright infringing material and threatens to bring legal action.[91]

The EU's Copyright Directive provides an exemption from copyright protection for text and data mining (TDM) purposes unless copyright holders have expressly reserved their rights

87 Federal Court Rules 'Big Data' Discrimination Studies Do Not Violate Federal Anti-Hacking Law, ACLU, 28 March 2020, https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti

88 Van Buren v United States (2021), United States Supreme court No. 19 -783, 03 June 2021, https://caselaw.findlaw.com/court/us-supreme-court/19-783.html

89 Van Buren is a Victory Against Overbroad Interpretations of the CFAA, and Protects Security Researchers, Aaron Mackay and Kurt Opsahl, Electronic Frontier Foundation, 03 June 2021, https://www.eff.org/deeplinks/2021/06/van-buren-victory-against-overbroad-interpretations-cfaa-protects-security

90 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 7; , Weiss, J. Petition for new exemption to section 1201 of the digital millennium copyright act: Exemption for security research pertaining to generative ai bias, June 2023, https://www.copyright.gov/1201/2024/petitions/proposed/New-Pet-Jonathan-Weiss.pdf. Grynbaum, M. M. and Mac, R. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. The New York Times, Dec 2023, https://www.nytimes.com/2023/12/27/business/media/new-york-times-openai-microsoft-lawsuit.html.

91 Midjourney Terms of Service, https://docs.midjourney.com/docs/terms-of-service.

(e.g. via opting out).[92] Developers have heavily relied on the TDM exemption in the development and training of foundation models. To support opting out, the EU's AI Act requires a sufficiently detailed summary of training data to be made available,[93] although this still places the onus on copyright holders to seek out the necessary information and use opt outs across developers. If the model produces outputs that include copyrighted material it is likely that the TDM exemption would apply, but there remains some uncertainty about whether liability could be placed on both the developer and the actor that created and used the prompt to produce the copyrighted work. This may cause additional legal uncertainty for external researchers conducting evaluations and red teaming related to copyright.

Data protection legislation such as the EU's General Data Protection Regulation poses less of a legal risk for researchers, as much foundation model research will not involve personal data. This is an important difference to social media research which generally involves the collection and analysis of the massive range of personal data generated through platform use. However, model evaluations that leak personal information from model training sets may expose researchers to legal risk.

Overall, this complex legal landscape has led to some unease amongst researchers about whether they are at risk of legal retaliation for research that breach terms of service.[94] This is a particular concern for researchers who are investigating research avenues that don't align with and threaten developers' interests.[95]

> *Policy option 3:*
> *Develop comprehensive safe harbour and responsible disclosure policies.*

92 Article 4, Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32001L0029.

93 AI Act Article 52c.1(a)

94 Massachusetts Institute for Technology, Safe Harbour for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/.

95 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861 page 10.

4.5 Model versioning and stability

> Most developers do not offer sufficiently transparent model versioning or model stability, including information about the usage of additional components and filters, which hinders reproducible research.

Transparent model versioning and model stability is essential to enable reproducible research on the exact same model.[96] While some developers do not offer these features, there are a number of good practices in the industry. Cohere's API documentation explains how to easily view specific model versioning,[97] whilst OpenAI provides a technical workaround. Google offers stable model versions of Gemini with clear model versioning and shares the date when the model will be discontinued.[98]

To ensure truly accurate model versioning and stability, developers need to be transparent about the use of additional components such as filters, which may be added to pre-process user inputs and outputs, and content moderation systems on deployed systems. Each deployed system could have many of these filters in place at any one time and they can change rapidly. For example, the image generative systems Dall-E[99] and Google Gemini[100] included a filter that added additional text to user prompts to request that the image outputs of people were diverse.

Policy option 4:
Provide clear and accurate model versioning and model stability.

96 Bucknall and Trager, Structured Access for Third-Party Research on Frontier AI Models, Centre for the Governance of AI, 31 October 2023, available at: https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models p.3.

97 For cloud deployment via AWS SageMaker, Cohere, March 2024. https://docs.cohere.com/docs/amazon-sagemaker-setup-guide

98 Google, May 2024, https://cloud.google.com/vertex-ai/generative-ai/docs/learn/model-versioning#stable-versions-available

99 J Baum and J Villasenor, Rendering misrepresentation: Diversity failures in AI image generation, Brookings, 17 April 2024, https://www.brookings.edu/articles/rendering-misrepresentation-diversity-failures-in-ai-image-generation/

100 P Raghavan, Gemini image generation got it wrong. We'll do better. Google Gemini, 23 February 2024, https://blog.google/products/gemini/gemini-image-generation-issue/

4.6 Levels of model and data access

> Model evaluations require access to conduct sampling and fine-tuning, which are available via many developers' APIs. Yet researchers often don't have full access to base models, model families, and components such as filters and content moderation systems. Other research may require deeper levels of access to models, internal data, and documentation, which are currently unavailable.

The research domain and objective will impact the level of model access and types of data required. For example, red teaming for trust and safety may simply require access to prompt a deployed model, whilst research on child safety may require access to training data and datasheets documentation.[101] Currently, researchers conducting model evaluations via APIs are able to conduct some basic sampling (prompting) and fine-tuning (further training of the model).[102] In addition, OpenAI also provides the probabilities and top five logits of its models, which are the values representing the likelihood that a token (e.g. letter, word, or pixel) will be selected to appear next in the model's output.[103] At the application level, researchers are able to conduct basic sampling of developer-deployed chatbots and other applications. Basic sampling is sufficient for evaluations that aim to identify model behaviour.[104]

At a minimum, Anderljung et al. suggest evaluations require access to base models, model families, the components of a deployed AI system, background information on the model, third-party data on the model's impacts, and the ability to fine-tune the model.[105] Access to the 'base model', the versions of the model that lack some safety mitigations such as fine-tuning, are necessary to understand intrinsic characteristics of the model and risks if safeguards are disabled.[106] In closed systems, this can be provided via an API rather than

101 Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, Stanford Internet Observatory, 20 December 2023, https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-20.pdf

102 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893, p. 6.

103 Ibid, p. 6.

104 Ibid, p. 2.

105 Anderljung, Thornton Smith, Joe O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 November 2023, https://arxiv.org/abs/2311.14711 p.8.; Bucknall and Trager, Structured Access for Third-Party Research on Frontier AI Models, Centre for the Governance of AI, 31 October 2023, https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models.

106 Research has revealed that researchers can identify the top model layer with access to logits. However no research has found it possible to reveal additional layers of the model. See Finlayson, Ren, Swayamdipta, Logits of API-Protected LLMs Leak Proprietary Information, 15 March 2024, https://arxiv.org/abs/2403.09539v2

disclosing model weights. Model families need to be disclosed to conduct research into scaling.[107]

However, other research areas may require a deeper level of system access which is currently unavailable. This may not be feasible given the level of access it requires to proprietary information. For example, Bucknall and Trager suggest that interpretability research will additionally require the ability to inspect and modify a model.[108] Inspection access to model internals, including parameters, activations and attention, and embeddings are increasingly understood as useful to understand model behaviour and predictions.[109] Some developers, such as Midjourney, enable API users to edit certain parameters.[110] Casper et al. suggest that external audits require white-box access to model weights, activations and gradients.

Beyond access to the model or system itself, researchers may require access to non-public company data and internal documentation relating to design and development processes, such as data cards,[111] internal evaluations, content moderation, and usage. In this regard, Casper et al. argue that researchers may require outside-the-box access to system training and deployment information, including methodology, code documentation, hyperparameters, deployment details, and internal evaluations.[112]

> Leading developers do not provide access to training data due to legal and reputational risks and due to competitive reasons.

Access to training data has been crucial for public interest research on open-source foundation models. Audits of the open-source dataset LAION-5B used to train certain

107 E.g. Birhane, Prabhu, Han and Boddeti, On Hate Scaling Laws For Data-Swamps, June 2023, available at: https://arxiv.org/abs/2306.13141; Anderljung, Thornton Smith, Joe O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 Nov ember 2023, https://arxiv.org/abs/2311.14711 p.8.

108 Bucknall and Trager, Structured Access for Third-Party Research on Frontier AI Models, Centre for the Governance of AI, 31 October 2023, available at: https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models, p. 1, 2.

109 Anderljung, Thornton Smith, Joe O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 November 2023, https://arxiv.org/abs/2311.14711, p.8.

110 Parameter List, Midjourney Documentation, https://docs.midjourney.com/docs/parameter-list

111 Pushkarna, Zaldivar, Kjartansson, Google Research, Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI, 03 April 2022, https://arxiv.org/abs/2204.01075.

112 Casper, Ezell, Siegmann, Kolt, Lynn Curtis, Bucknall, Haupt, Wei, Scheurer, Hobbhahn, Sharkey, Krishna, von Hagen, Alberti, Chan, Sun, Gerovitch, Bau, Tegmark, Kreuger, Hadield-Menell, Black-Box Access is Insufficient for Rigorous AI Audits, 25 January 2024, https://arxiv.org/abs/2401.14446

foundation models have revealed racist[113] and child sexual abuse content.[114] Despite calls for transparency, it is unlikely that many of the closed developers will release their training data given the range of legal and reputational risks that arise from the possibility of problematic, harmful, and illegal content in datasets, and for competitive reasons.[115] For example, OpenAI,[116] Google,[117] and Midjourney,[118] are subject to ongoing legal challenges which allege that training data was obtained in violation of copyright rules. The presence of illegal content within training data, such as child sexual abuse and terrorism content, could also open developers up for liability under many jurisdictions' criminal and online safety laws.

At a minimum, researchers argue that developers should provide sufficient documentation to understand the data provenance and processing decisions,[119] such as datasheets for datasets[120] As a potentially helpful starting point, the EU's AI Act will require developers of so-called "general-purpose AI models" to make publicly available a sufficiently detailed summary of the content used for training, and the European Commission will develop a reporting template.[121]

> *Policy option 5:*
> *Provide researchers with sufficient levels of model and data access via structured researcher access program.*

113 Birhane, Prabhu, Han and Boddeti, On Hate Scaling Laws For Data-Swamps, June 2023, available at: https://arxiv.org/abs/2306.13141

114 Thiel, Identifying and Eliminating CSAM in Generative ML Training Data and Models, Stanford Internet Observatory, 20 December 2023, https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-20.pdf

115 OpenAI, GPT-4 Technical Report, March 2024, https://arxiv.org/pdf/2303.08774.

116 The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work, New York Times, https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html; David, The Intercept, Raw Story, and AlterNet sue OpenAI and Microsoft, 28 February 2024, https://www.theverge.com/2024/2/28/24085973/intercept-raw-story-alternet-openai-lawsuit-copyright#

117 Milićević, France fines Google €250 million for copyright infringement, 20 March 2024, https://dig.watch/updates/france-fines-google-e250-million-for-copyright-infringement

118 Vincent, AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, 16 January 2023, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

119 Longpre, Mahari, Chen, Obeng-Marnu, Sileo, Brannon, Muennighoff, Khazam, Kabbara, Perisetla, X A Wu, Shippole, Bollacker, T Wu, Villa, Pentland, Hooker, The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI, 4 November 2023, https://arxiv.org/pdf/2310.16787; Bommasani, Klyman, Longpre, Xiong, Kapoor, Maslej, Narayanan, Liang, Foundation Model Transparency Reports, 26 February 2024, https://arxiv.org/pdf/2402.16268.pdf; https://arxiv.org/abs/2310.12941; Anderljung, Thornton Smith, Joe O'Brien, Soder, Bucknall, Bluenke, Schuett, Trager, Strahm, and Chowdhury, Towards Publicly Accountable Frontier LLMs, 15 November 2023, https://arxiv.org/abs/2311.14711, p.8.

120 Gebru, Morgenstern, Vecchione, Wortman Vaughan, Wallach, Daumé III, Crawford, Datasheets for Datasets, 01 December 2021, https://arxiv.org/abs/1803.09010.

121 AI Act Article 52c.1(d)

4.7 Access to usage data

> Developers do not share any information about usage trends. Researchers are interested in accessing this information to inform further research and policymaking. This information can only be shared in a manner that respects users' strong expectation of privacy, given the private nature of interactions with foundation models.

Developers are logging and processing user behavioural data to identify malicious actors and enforce usage policies, including user prompts and usage rates. OpenAI collects usage data and processes it in accordance with its data usage policy.[122] Researchers are calling for developers to monitor, taxonomise, and share information about how end-users are using systems to inform research and policymaking.[123]

Sharing individual user behavioural data raises privacy concerns. Unlike social media platforms (where user data is generated and shared on publicly accessible platforms), users generally interact with foundation models and downstream applications in private environments and may ask questions containing personal or commercially sensitive information. Unless a user publishes the prompt and response elsewhere, they will have a strong expectation that this data will remain private. Indeed, many developers' privacy policies assure users that this will be the case and provide the option for users to opt-out of contributing their usage data to further model training.

If researchers are interested in collecting or accessing usage data, research will need to be conducted under strict privacy controls and researchers may need to be subject to vetting or accreditation. Lessons should be drawn from social media researchers' best practices, including the Digital Services Act access to data provisions[124] and the European Data Media Observatory's Code of Conduct on Researcher Access to Data.[125] Third party data collection, such as crowdsourcing approaches, are important because researchers do not need to rely on developers providing the necessary information or relying on access to a developer's

122 OpenAI, Europe Privacy Policy, 15 December 2023, https://openai.com/policies/privacy-policy/.

123 Red Teaming Isn't Enough, G. Nicholas, Foreign Policy, 08 July 2024, https://foreignpolicy.com/2024/07/08/artificial-intelligence-ai-election-misinformation-technology-risks/?tpcc=recirc_latest062921; Caliskan and Lum, Effective AI regulation requires understanding general-purpose AI, 29 January 2024, https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/

124 Article 40 Digital Services Act.

125 European Digital Media Observatory, Institute for Data, Democracy & Politics, The George Washington University, Report of the Europeam Media Observatory's Working Group on Platform-to-Researcher Data Access, 31 May 2022, https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

proprietary system.[126] This is particularly important in this context, as data about usage is not currently made available. Even if developers begin to share this data, third party approaches will be necessary for verification.[127] Indeed, in the social media context, third party data collection has proved to be more reliable, contemporaneous, and accurate than information provided via a structured researcher API.[128]

Policy option 6:
Develop or enable crowdsourced data collection approaches for foundation model research, including to collect usage information.

4.8 Transparency reports and documentation

Developers are not sufficiently transparent in supporting documentation, particularly in relation to environmental and labour practices.

Documentation-based approaches complement evaluations by providing broader contextual information. The Foundation Model Transparency Index assessed foundation model developers against 100 transparency indicators categorised across three domains, 1) upstream resources used to build a model, 2) model properties and evaluations, and 3) downstream use and impacts.[129] Unsurprisingly, the initial assessment found that closed foundation model developers are less transparent than the open developers (Meta, Hugging Face and Stability.AI).[130] However, all of the developers scored particularly poorly on labour, usage statistics, and downstream impact. This current inability to access such information poses a challenge for social impacts research.

Policy option 7:
Publish transparency reports.

126 Nicholas, Red Teaming Isn't Enough, Foreign Policy, 08 July 2024, https://foreignpolicy.com/2024/07/08/artificial-intelligence-ai-election-misinformation-technology-risks/?tpcc=recirc_latest062921; Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861.

127 Abdo, Krishnan, Krent, Falcón and Woods, A Safe Harbor for Platform Research, Knight First Amendment Institute, 19 January 2023, https://knightcolumbia.org/content/a-safe-harbor-for-platform-research

128 Ibid.

129 Bommasani, Klyman, Longpre, Xiong, Kapoor, Maslej, Narayanan, Liang, Foundation Model Transparency Reports, 26 February 2024, https://arxiv.org/pdf/2402.16268.pdf

130 The Foundation Model Transparency Index, Center for Research on Foundation Models, https://crfm.stanford.edu/fmti/. This has been criticised for misleadingly conflating commercial documentation with transparency, creating an incentive to optimise scores rather than meaningfully improve transparency, and for factual errors. See Lambert, Gyges, Biderman, Skowron, How the Foundation Model Transparency Index Distorts Transparency, 26 October 2023, https://blog.eleuther.ai/fmti-critique/.

4.9 Developer control over access

> Developers are gatekeepers of access programs, which enable them to prioritise research that aligns with internal priorities or is less commercially threatening. Compounding concerns, selection processes may not be sufficiently resourced or transparent enough to promote trust.

The exclusive developer control over the majority of researcher access initiatives means developers are free to act as gatekeepers,[131] deciding which researchers and what research aims should benefit from what level of access.[132] This has caused friction with the UK AISI, with the organisation unable to get pre-release access to models, in breach of voluntary commitments. In relation to industry-led access programs, many application processes require researchers to disclose their research plan and profile.[133] In addition, access to some developers' commercial API also requires an explanation about the use of the API. While these are valid criteria to screen researchers, they do provide an opportunity for industry to deprioritize legitimate requests that don't fit the developer's priorities.

For example, Inflection chose to involve external red teamers in the domain of bioengineering, as the developers considered it a high-risk domain.[134] OpenAI acknowledged that its pre-deployment red teaming of GPT-4 was biased towards English-speaking researchers affiliated to academia in the US, Canada, and UK,[135] but has since recruited to improve geographic and domain diversity through its red teaming network.[136] Both Cohere and OpenAI's researcher access to API programs describe the research areas they are most interested in which may influence researchers' work.

Compounding these concerns, the selection processes for several industry programs are not sufficiently transparent and may be under-resourced. As a result of this lack of trans-

131 AI Now Institute, Algorithmic Accountability: Moving Beyond Audits, 11 April 2023, https://ainowinstitute.org/publication/algorithmic-accountability#weak-policy-response

132 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861, page 8, Longpre, Kapoor, Klyman, Ramaswani, Bom masani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page.

133 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861 page 8.

134 Written Testimony of Dario Amodei, Ph.D. Co-Founder and CEO, Anthropic, Judiciary Committee Subcommittee on Privacy, Technology, and the Law United States Senate, 25 July 2023, https://www.judiciary.senate.gov/imo/media/doc/2023-07-26_-_testimony_-_amodei.pdf

135 https://cdn.openai.com/papers/gpt-4-system-card.pdf

136 https://openai.com/blog/red-teaming-network

parency, Longpre et al., found, for example, that there was a "strong impression that access to OpenAI employees improves access to their programs."[137] Longpre et al. also found that developers may have backlogs partly due to dedicating "few resources" to the selection process.[138]

Policy option 8:
Create an independently mediated structured researcher access program.

137 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893. page 4.

138 Ibid, page 4.

5.

Recommendations

Section 5: Recommendations

Recommendation 1: Provide researchers with subsidised API access.

To promote third-party discovery of vulnerabilities, developers could provide researchers with subsidised API access with sufficiently high rate-limits to conduct automated evaluations. This should be allocated to researchers across a wide range of research domains and subject matter expertise, and include support for researchers conducting longer-term research rather than discrete projects (e.g. monitoring). This could be allocated by developers via industry researcher access programs or by independent reviewers[139] at expert public bodies such as the US's NAIRR in the NSF or AISIs.

Recommendation 2: Provide transparent and effective content moderation and appeals processes, with fast-track appeals for researchers.

As Longpre et al. advocate, developers could provide a transparent and well-resourced appeals process with a public commitment to restore the accounts of good faith external researchers.[140] To enable user appeals, developers should provide justifications for enforcement actions taken. Developers need to sufficiently resource appeals processes conducted by independent reviewers, offer a fast-track process for researchers (e.g. upon submission of research organisation affiliation), make decision criteria and outcomes visible to the wider community, and guarantee response times.[141]

Longpre et al. offer one 'Technical Safe Harbor' approach in which researchers pre-register their profile and research plan in advance so that developers can easily cross-reference and

139 Massachusetts Institute for Technology, Safe Harbour for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/

140 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 8.

141 Ibid, page 9; Massachusetts Institute for Technology, Safe Harbour for Independent AI Evaluation, https://sites.mit.edu/ai-safe-harbor/.

review flagged accounts.[142] This would allow the developers to review the accounts behaviour, against the researchers' profile and plan including whether policy violations are in line with the research plan and have been reported under a responsible disclosure policy.

Recommendation 3:
Develop comprehensive safe harbour and responsible disclosure policies.

As a constellation of AI researchers propose, developers could provide a clear and explicit safe harbour policy for good faith security and safety researchers.[143] To ensure that safe harbours are comprehensive, further research into the legal risks of security and safety researchers would be beneficial. This research should build on the history and experience of hackers, social media researchers, and other independent technology and security researchers, mindful of the differing contexts and access involved.

To promote a uniform level of protection, existing fora such as the EU-US Trade and Technology Council could be used to develop a harmonised legal safe harbour, and responsible disclosure policy, which could also lead to a standardisation requests to, for example, the National Institute of Standards and Technology (NIST).[144] Safety researchers would also benefit from policy guidance on whether safety evaluations and third-party data collection methods would violate laws. If deemed appropriate, the US Department of Justice and EU Member States could also publish prosecutorial safe harbour policies for good faith safety research in line with recent policies.

Recommendation 4:
Provide clear and accurate model versioning and stability.

Developers need to provide clear and accurate model versioning and stability. This should also extend to the transparent and stable use of additional components on deployed systems, such as filters and content moderation systems.

142 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 18.

143 Ibid; A Safe Harbour for Independent AI Evaluation, Massachusetts Institute for Technology, https://sites.mit.edu/ai-safe-harbor/.

144 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 7.

Recommendation 5:
Provide researchers with sufficient levels of model and data access via a structured access program.

Research that requires deeper levels of model access and access to non-public internal data and documentation could be mediated through a structured access program subject to a vetting or accreditation procedure to ensure sufficient levels of confidentiality can be maintained. This could occur via a secure environment, such as a researcher-specific API, a query-based system, or facilitating on-site terminal access.[145] Model access via a structured researcher program will only be appropriate for some research avenues. Researchers interested in evaluating the safety or security properties of a deployed foundation model should instead access the model through its widely used commercial API or public interface. Otherwise, there is a risk a company may make a slightly different model available through the research API compared to the publicly available interfaces.

At a minimum, Bucknall and Trager argue a research API and accompanying documentation should include: (1) Increased transparency regarding model information, such as model versioning, size, fine-tuning processes, and information about the datasets used in pre-training. (2) Output logits and the ability to choose from and modify different sampling algorithms. (3) Version stability and back-compatibility features to enable reproducible research on the same model. (4) Ability to fine-tune the model. (5) Access to model families to understand how the models systematically differ. It will be necessary for researchers and industry to co-design such an API to ensure it accurately captures the existing and possible schema.

Trask et al. build on the API approach to propose a more flexible and open-ended approach to external audits.[146] This would entail several steps: auditors accessing an API provided by a model developer with access to a mock AI system and mock data, preparing and submitting audit code to be executed on the real system and data by the model developer; and subsequently downloading the audit results. One similar proposal is to create a National Deep Inference Facility in the National Science Foundation.[147] The facility would provide the necessary software-hardware infrastructure to enable researchers to conduct query-based audits of foundation models.

145 OpenMined, How to Audit an AI Model Owned by Someone Else (Part 1), 01 July 2023, https://blog.openmined.org/ai-audit-part-1/

146 OpenMined, How to Audit an AI Model Owned by Someone Else (Part 1), 01 July 2023, https://blog.openmined.org/ai-audit-part-1/

147 Bau, A National Deep Inference Facility, The Visible Net, 04 July 2023, https://thevisible.net/posts/003-national-deep-inference-facility/

Finally, on-site access may be required to facilitate out-of-the-box audits.[148] This involves external auditors accessing the relevant models, systems, and data which are stored in a secure environment.[149] It is possible that these audits will need to be conducted by certified and regulated auditors subject to strict conduct and confidentiality agreements.[150]

Non-public data that may be relevant to request could include internal documentation relating to design and development processes, such as data cards,[151] internal evaluations, content moderation, and usage. If training data is not available, datasheets[152] could be shared and, at minimum, documentation should include clear listing of the primary data collections or sets utilised throughout training, such as large private or public databases, along with narrative explanations of other data sources used and the data processing conducted, as good practice in line with the EU's AI Act.[153]

Recommendation 6: Develop or enable crowdsourced data collection approaches for foundation model research, including to collect usage information.

Developers and external researchers could explore the use of crowdsourced data collection to support research. This could include opt-in data donation and web scraping tools to collect usage data. Researchers working with usage information will need to respect data protection regulation and norms, and should build on social media researchers' best practices.

For example, developers of downstream applications could enable users to opt-in to donate usage data, either at a granular or aggregated level, for research purposes. Already, Midjourney's privacy policy states it uses personal data to identify usage trends but this information is not made available to external actors.[154]

148 Casper, Ezell, Siegmann, Kolt, Lynn Curtis, Bucknall, Haupt, Wei, Scheurer, Hobbhahn, Sharkey, Krishna, von Hagen, Alberti, Chan, Sun, Gerovitch, Bau, Tegmark, Kreuger, Hadield-Menell, Black-Box Access is Insufficient for Rigorous AI Audits, 25 January 2024, https://arxiv.org/abs/2401.14446 page 11.

149 How to Audit an AI Model Owned by Someone Else (Part 1), OpenMined, 01 July 2023, https://blog.openmined.org/ai-audit-part-1/

150 Casper, Ezell, Siegmann, Kolt, Lynn Curtis, Bucknall, Haupt, Wei, Scheurer, Hobbhahn, Sharkey, Krishna, von Hagen, Alberti, Chan, Sun, Gerovitch, Bau, Tegmark, Kreuger, Hadield-Menell, Black-Box Access is Insufficient for Rigorous AI Audits, 25 January 2024, https://arxiv.org/abs/2401.14446 page 12.

151 Pushkarna, Zaldivar, Kjartansson, Google Research, Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI, 03 April 2022, https://arxiv.org/abs/2204.01075.

152 Datasheets for Datasets, Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford, 01 December 2021, https://arxiv.org/abs/1803.09010.

153 Open Future and Mozilla Foundation, Sufficiently detailed? A proposal for implementing the the AI Act's training data transparency requirements for GPAI, June 2024, https://openfuture.eu/wp-content/uploads/2024/06/240618AIAtransparency_template_requirements-2.pdf.

154 Privacy Policy, Midjourney Documentation, https://docs.midjourney.com/docs/privacy-policy

Alternatively, third-party data collection tools could be used. These are an important method for researchers to gather information outside of company-controlled interfaces.[155] For example, a web-browser based scraping tool (similar to The Markup's Citizen Browser[156]) could be developed to enable users to consensually donate prompts and responses, with the potential to offer granular consent to specific research projects. Alternatively, surveys or user interviews could be used, although this would likely miss malicious or embarrassing use cases.[157]

Recommendation 7: Publish transparency reports

Developers could publish regular transparency reports that cover content moderation and appeals processes and data, and usage trends (if collected), environmental impacts, and labour impacts. Specific information about the prevalence and reasons for content moderation against registered researchers would enable the wider research community to apply pressure against attempts to restrict legitimate research. Given the potential environmental and labour impacts of foundation models, information concerning energy costs, carbon emissions, and labour in the supply chain could be published as a matter of due diligence. [158]

These reports could be inspired by the Foundation Model Transparency Index[159], although this has been criticised.[160] Alternatively, inspiration could be taken from regular transparency reporting obligations on social media platforms, as mandated by the DSA, for example.

155 Ojewale, Steed, Vecchione, Birhane, Raji, Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling, 27 February 2024, https://arxiv.org/abs/2402.17861 page 8.

156 The Markup, Launching Citizen Browser, 05 January 2025, https://themarkup.org/newsletter/citizen-browser/launching-citizen-browser

157 Caliskan and Lum, Effective AI regulation requires understanding general-purpose AI, 29 January 2024, https://www.brookings.edu/articles/effective-ai-regulation-requires-understanding-general-purpose-ai/

158 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 9.

159 Bommasani, Klyman, Longpre, Xiong, Kapoor, Maslej, Narayanan, Liang, Foundation Model Transparency Reports, 26 February 2024, https://arxiv.org/pdf/2402.16268.pdf

160 Criticised for misleadingly conflating commercial documentation with transparency, creating an incentive to optimise score rather than meaningfully improve transparency, and for factual errors, see Lambert, Gyges, Biderman, Skowron, How the Foundation Model Transparency Index Distorts Transparency, 26 October 2023, https://blog.eleuther.ai/fmti-critique/.

Recommendation 8:
Create an independently mediated structured researcher access program.

Developers should not be the final arbiter to decide which researcher or research proposals merit privileged access to their models, including pre-release models or via subsidised API access, and non-public data and documentation. Instead, structured access may need to be delegated to an independent mechanism. This mechanism could be involved in the selection and vetting of researchers conducting pre-release and post-deployment research, allocate subsidised API access, and facilitate structured access by authorising access to models, data, and documentation via a specific researcher API, sandbox or query system. The program should be flexible and iterative, considering the experience of researchers to re-design and improve its mechanisms.[161]

This could be done on a voluntary basis through existing publicly funded bodies such as the NAIRR Pilot or the AISI, or via an industry-funded organisation.[162] Liang et al. propose a foundation model review board to mediate the selection of external research proposals.[163] This body aims to facilitate the process of developers releasing models to external researchers. Researchers would need to submit the goals of the research, type of access required, an ethics statement, and any related proposals. The board would review research proposals and make recommendations to developers on which research proposals to select. Ultimately, the developer maintains control about whether to approve, reject, or defer the proposal. If the proposal is accepted, the developer will release the desired assets to the researcher via an API. The board would need to be representative of the broader research community such as academia, industry, civil society, and impacted groups.[164]

However, given some reticence to work with the UK AISI, it may be necessary to legally mandate external structured access, both to AISIs and authorised third party researchers. Article 40(4) of the EU's Digital Service Act on structured researcher access provides a similar model but leaves less autonomy to developers to reject a proposal, only allowing rejections if the developer doesn't have the data or giving access to the data would lead to significant vulnerabilities in the security of their service or the protection of confidential information. This would need to be on a legislative footing and require a regulatory body.

161 Van Drunen, Noroozian, How to design data access for researchers: a legal and software development perspective, https://www.sciencedirect.com/science/article/pii/S026736492400013X

162 Social media platform companies have voluntarily agreed to set up such an industry-funded body under the EU's Code of Conduct on Disinformation.

163 Liang, Bommasani, Creel Reich, The Time Is Now to Develop Community Norms for the Release of Foundation Models, Stanford University Human-Centred AI, 17 May 2022, https://hai.stanford.edu/news/time-now-develop-community-norms-release-foundation-models

164 Ibid.

In order to facilitate the vetting tasks of this regulator the Digital Services Act foresees a potential role for an 'independent advisory' mechanism to assist in vetting applications. This independent mechanism could be better positioned than a national regulator to assess the scientific merits – and societal benefits – of a potential research project.

This mechanism would require sufficient resources and be staffed with a diverse group of experts across research domains and testing methods, including evaluations, red teaming, and social impact work. Given the limited number of experts, it would likely need to be a feature of existing specialised bodies. In the US, the National Science Foundation may be most appropriate since it already disperses compute resources via the NAIRR Pilot based on researcher applications, and the US Artificial Intelligence Safety Institute Consortium (AISIC) could assist as an independent advisory body. In the EU, the EU AI Office will have the relevant expertise to assess applications, but will require additional resourcing given its existing responsibilities under the AI Act. The UK AISI – and soon the US AISI – also has expertise in conducting and assessing evaluations of models, recruiting external experts, and is familiar with many developers. Longpre et al. propose that multiple organisations be involved in allocating model access, and each allocated a specific API key that the organisation can authorise for use amongst its own network of researchers whilst retaining responsibility for misuse.[165]

It would be beneficial to conduct a multi-stakeholder consultation to agree on the institutional design, appropriate levels of system access, infrastructural design, rules around length and retention of access, confidentiality, and publication of research.[166] Building on the experience of social media researchers, there should be the possibility for access to be provided on an ongoing basis to specific researchers for specific research (e.g. monitoring) rather than on the basis of discrete project proposals.

165 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 8.

166 Lurie, Comparing Platform Research API Requirements, Tech Policy Press, 22 March 2023, https://www.techpolicy.press/comparing-platform-research-api-requirements/

Conclusion

As the use and integration of foundation models in a wide range of products and services accelerates, the impacts of potential safety and security risks will materialise at increasing rates. Alongside foundation model developers conducting tests and implementing internal safeguards, and emerging frameworks of regulatory due diligence obligations, external research plays an important complement in uncovering potential and existing harms and vulnerabilities, proposing fixes, and holding developers to account.

Yet, as this report demonstrates, public-interest researchers studying foundation models lack access, information, resources, and legal certainty. In this respect, leading developers are effectively operating as gatekeepers of AI safety. To advance safety and security in the AI ecosystem, it is time for developers and governments to work with external researchers and civil society and chart the path for a flourishing public-interest research ecosystem.

Annexes

Annex 1 – Mapping of pre-release industry-led external researcher access programs.

ANNEX 1				
Company	**Program information**	**Access**	**External expertise sought**	**Selection process**
OpenAI[167]	50 external researchers were given pre-release access to GPT-4 to conduct red teaming and evaluations.	Early versions of GPT-4 and models with in-development mitigations.	Fairness, alignment, industry trust and safety, dis-/misinformation, chemistry, bio-risk, cybersecurity, nuclear risks, economics, human-computer interaction (HCI), law, education, and healthcare.	Selected researchers based on prior research or experience. They primarily had significant higher education or industry experience and ties to English-speaking Western countries.[168]
OpenAI[169]	In December 2023, OpenAI formalised the Red Teaming Network. It develops taxonomies of risk and conducts evaluations. This is compensated.	Access at various stages in model and product development lifecycle, occasionally deeper levels of system access including to base models.	Cognitive science, biology, computer science, political science, persuasion, anthropology, HCI, alignment, healthcare, child safety, finance, biometrics, political use, chemistry, physics, steganography, psychology, economics, sociology, fairness and bias, education, law, cybersecurity, mis/disinformation, privacy, languages, and linguistics	Published a call for applications. The selection process stated a concern for geographic and domain diversity. There is no public information about successful applicants.

167 OpenAI, GPT-4 Technical Report, 04 March 2024, https://arxiv.org/pdf/2303.08774.pdf ; https://cdn.openai.com/papers/gpt-4-system-card.pdf

168 OpenAI, GPT-4 System Card, 23 March 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf

169 OpenAI, Red Teaming Network, https://openai.com/blog/red-teaming-network

ANNEX 1				
Company	**Program information**	**Access**	**External expertise sought**	**Selection process**
Anthropic[170]	Gryphon Scientific was given pre-release access to models to conduct red teaming.	Pre-release models.	Biosecurity.	Unclear.[171]
Anthropic[172]	External domain experts are commissioned to conduct red teaming throughout the AI life cycle in three areas: trust and safety, national security, and non-American English languages and contexts.	Access at various stages in model and product development lifecycle.	Policy Vulnerability Testing for Trust & Safety risks. Worked with Thorn on child safety, ISD on election integrity and Global Project Against Hate and Extremism on radicalisation. Frontier threats testing for national security risks, focuses on Chemical, Biological, Radiological and Nuclear, cybersecurity and autonomous AI risks. Multilingual and multicultural red teaming with public sector agencies. Worked with Singapore's Infocomm Media Development Authority and AI Verify Foundation to red team across English, Tamil, Mandarin and Malay.	Unclear. Works with subject matter experts.

170 Anthropic, Frontier Threats Red Teaming for AI Safety, 26 July 2023, https://www.anthropic.com/news/frontier-threats-red-teaming-for-ai-safety#entry:146918@1:url

171 Gryphon Scientific is an active research institute on foundation models. It is a member of the US's AISIC and a technical partner of the UK's AISI.

172 Anthropic, Challenges in red teaming AI systems, 12 June 2024, https://www.anthropic.com/news/challenges-in-red-teaming-ai-systems

ANNEX 1				
Company	**Program information**	**Access**	**External expertise sought**	**Selection process**
Inflection[173]	External experts are commissioned to conduct pre-release red teaming of models.	Pre-release models.	Previous collaborations with mental health professionals, and biosecurity experts. They are recruiting particularly across chemical, biological, radiological, and nuclear risks.	Unclear.

173 Inflection, Our policy on frontier safety, 30 October 2023, https://inflection.ai/frontier-safety

Annex 2 – Mapping of researcher access to API programs.

ANNEX 2				
Company	**Research API program**	**Subsidy**	**Research areas in scope**	**Selection process**
Industry-led programs				
OpenAI[174]	Yes.	Yes, via API credits.	Alignment, fairness and representation, societal impact, interdisciplinary researcher, interpretability and transparency, misuse potential and robustness.	For researchers with limited financial and institutional resources. Applications must include research use case and researcher profile. Applications are processed within 4 – 6 weeks and unsuccessful applications do not receive a response. There is a "strong impression that access to OpenAI employees improves access to their programs."[175]
Google	No.			

174 OpenAI, Researcher Access Program, https://openai.com/form/researcher-access-program

175 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 page 4.

ANNEX 2				
Company	**Research API program**	**Subsidy**	**Research areas in scope**	**Selection process**
Anthropic[176]	Yes, provides $1000 API credits to standard model suite. Does not receive an exemption from usage policy and is unable to conduct fine-tuning.	Yes.	Safety and alignment.	Applications must include a research plan and researcher profile. Applications are evaluated on the first Monday of each month.
Inflection	No.			
Midjourney	No. Access is only available at the application layer via Midjourney Bot. Its usage policy specifically prohibits automated interactions.[177]			
Cohere[178]	Yes, provides grants and model access.	Yes.	Research into good application of LLMs (e.g. climate science and content moderation), LLM safety (bias, explainability, hallucinations, synthetic data quality, toxicity, adversarial testing), LLM capabilities, multilingual capabilities and value alignment.	Aimed at researchers from academia and third-party institutions that want to conduct research with the goal of releasing a "peer-reviewed scientific artifact." Applications must include research or use case and goals, researcher profile, and models included in the study. It does not provide an application review time frame.

176 Anthropic, What is the external researcher access program?, https://support.anthropic.com/en/articles/9125743-how-can-i-access-the-claude-api-for-alignment-research-purposes; Anthropic, Prioritising research on risks posed by AI, https://www.anthropic.com/uk-government-internal-ai-safety-policy-response/prioritising-research-on-risks-posed-by-ai;

177 Midjourney, Community Guidelines, https://docs.midjourney.com/docs/community-guidelines

178 Cohere, Cohere for AI Research Grants, 11 July 2023, https://txt.cohere.com/c4ai-research-grants/; https://cohere.com/blog/granting-access

ANNEX 2				
Company	**Research API program**	**Subsidy**	**Research areas in scope**	**Selection process**
National AI Research Resource Pilot[179]				
OpenAI	Yes, via NAIRR Pilot.	Yes, donated up to $1.25 million in credits for model access	$1 million for AI safety, evaluations, and societal impacts. $250,000 to support applied research at HBCUs.	The NAIRR supports US-based researchers with the computational resources necessary to conduct evaluations and research across AI systems. Researchers must apply via the National Science Foundation, including the research use case and researcher profile.
Microsoft	Yes, via NAIRR Pilot.	Yes, donated $20 million in compute credits on Microsoft Azure, along and access to models via Azure OpenAI (e.g. GPT-4).	Trustworthy and responsible AI applications, including fairness, accuracy, reliability, transparency, privacy, and security.	
Anthropic	Yes, via NAIRR Pilot.	Yes, for ten researchers.	Climate change related projects. Evaluations or research on Claude are out of scope.	

179 National Artificial Intelligence Research Resource https://new.nsf.gov/focus-areas/artificial-intelligence/nairr. We have only included companies that provide access to closed foundation models.

ANNEX 2				
Company	**Research API program**	**Subsidy**	**Research areas in scope**	**Selection process**
AI Village's Generative AI Red Team Event, DEF CON 2023[180]				
OpenAI, Google, Cohere, Anthropic[181]	Testing APIs provided via secure laptops. APIs had high rate-limits and technically exempt from moderation processes.	Yes, access to models via APIs were donated for the event.	Model safety (Factuality, Bias, Misdirection) and Cybersecurity.	Open to the public. Involved 2,244 attendees, including safety and security researchers, community groups, policy-oriented non-profits, and interested government parties.

180 Catell, AI Village, Generative Red Team Recap, 12 October 2023, https://aivillage.org/defcon%2031/generative-recap/; Humane Intelligence, Generative AI Red Teaming Challenge, https://www.humane-intelligence.org/grt

181 We have only included the companies that provided access to closed foundation models. Other industry partners were Hugging Face, Stability.ai and NVIDIA.

Annex 3 – Mapping of enforcement processes

ANNEX 3				
Company & System	**Enforcement processes for API and Application users**	**Suspensions against researcher accounts**	**Justification given for specific case**	**Appeals process**
OpenAI (GPT-4)	Human reviewers, machine learning & rule-based classifier detection systems to monitor user behaviour and identify violating content.[182] May result in warnings, account suspension or deletion.	Yes.[183]	No.	Yes clear violations of ChatGPT includes a link to an appeals process.[184]
Google (Gemini)	Yes. Has processes to prevent core harms and states they "may review content." [185] No information about specific processes involved. Access to content provided by API may be restricted, limited or filtered.	Unclear.	Yes, when prompt or query is blocked or deemed violative.[186]	Unclear.

182 OpenAI, GPT-4 Technical Report https://arxiv.org/pdf/2303.08774.pdf, p. 62, 66.

183 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 p5.

184 Bommasani, Scores for OpenAI on 2023 Foundation Model Transparency Index, https://github.com/stanford-crfm/fmti/blob/main/scoring/OpenAI%202023%20FMTI%20Scores.pdf, p.17.

185 Google AI for Developers, Gemini API Safety Settings, https://ai.google.dev/docs/safety_setting_gemini; https://developers.google.com/terms#a_api_prohibitions

186 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 p.18.

ANNEX 3				
Company & System	**Enforcement processes for API and Application users**	**Suspensions against researcher accounts**	**Justification given for specific case**	**Appeals process**
Anthropic[187] (Claude 2)	Yes. Will implement "detections and monitoring" but no information about specific processes involved. May result in warnings, account throttling, suspension or deletion.	Yes.[188]	Unclear.	Unclear

187 Anthropic, Usage Policy, 06 June 2024, https://www.anthropic.com/legal/aup; Bommasani, Scores for Antrhopic on 2023 Foundation Model Transparency Index, https://github.com/stanford-crfm/fmti/blob/main/scoring/Anthropic%202023%20FMTI%20Scores.pdf, point 75.

188 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893., p5.

ANNEX 3				
Company & System	**Enforcement processes for API and Application users**	**Suspensions against researcher accounts**	**Justification given for specific case**	**Appeals process**
Inflection (Pi)	Yes. Has human safety team and "tripwire" systems to identify behaviour to undermine safety or use models for inappropriate or harmful purposes, including behavioural patterns associated with systematic efforts. [189] It is also "experimenting" with using large language models to identify misuse on its platform. May result in warnings, account suspension or deletion.	Yes.[190]	Yes when prompt or query is blocked or deemed violative.[191]	Yes.[192]

189 Inflection, Our policy on frontier safety, 30 October 2023, https://inflection.ai/frontier-safety

190 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 p5.

191 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 p.18.

192 Pi Support Desk, Understanding Account Suspension, https://heypisupport.zendesk.com/hc/en-us/articles/17791183959437-Understanding-Account-Suspension-Why-was-my-account-suspended

ANNEX 3				
Company & System	**Enforcement processes for API and Application users**	**Suspensions against researcher accounts**	**Justification given for specific case**	**Appeals process**
Mid-journey[193] (v6)	Yes. Collects username, text and image prompt inputs, public chats, IP address to monitor usage of its service.[194] Does not share information about the specific processes involved. May result in account suspension or deletion.	Yes.[195] Repeatedly deleted accounts of researchers evaluating propensity to produce copyright infringing material.[196]	Unclear.	Unclear.[197]
Cohere (Command)	Unclear. Requires API users to report violative usage and content within 24 hours.[198]	Unclear.	Unclear.	Unclear.

193 Midjourney, Terms of Service, https://docs.midjourney.com/docs/terms-of-service

194 Midjourney, Privacy Policy, https://docs.midjourney.com/docs/privacy-policy

195 Longpre, Kapoor, Klyman, Ramaswani, Bommasani, Blili-Hamelin, Huang, Skowron, Xin, Kotha, Zeng, Shi, Yang, Southen, Robey, Chao, Yang, Jia, Kang, Pentland, Narayanan, Liang, Henderson, A Safe Harbor for AI Evaluation and Red Teaming, 05 March 2024, https://arxiv.org/abs/2403.04893 p5.

196 Marcus, Southen, IEEE Spectrum,Generative AI has a Visual Plagiarism Problem, 06 January 2024, https://spectrum.ieee.org/midjourney-copyright

197 Longpre et al. sugxgest there is an appeals process, but we were unable to verify this.

198 Cohere, Model Limitations, https://docs.cohere.com/docs/model-limitations

Annex 4 – Mapping of vulnerability reporting processes.

ANNEX 4				
Company & System	**Reporting tools**		**Legal safe harbour**	
	Security	**Safety**	**Security**	**Safety**
OpenAI (GPT-4)	Yes. Provides a bug bounty program with financial rewards.[199]	Yes. Form for 'Model behaviour feedback'.[200]	Yes, including the DMCA and CFAA. For authorised good faith research that complies with policy.[201]	Unclear. Does not include an accompanying policy, legal safe harbour statement, or rewards.
Google (Gemini)	Yes. Provides a bug bounty program for security and safety researchers to report vulnerabilities.[202] This includes training data extraction, manipulating models, model theft, and prompt attacks (except those that generate violative content because there is a dedicated reporting channel for this).[203] Prompt attacks that generate violative content or including copyright infringing content are out of scope and each have dedicated reporting channels.[204]		Unclear. The bug bounty program does not include a legal safe harbour statement and requires testing to "not violate any law, or disrupt or compromise any data that is not your own."	

199 Vulnerability Disclosure Policy, OpenAI https://openai.com/policies/coordinated-vulnerability-disclosure-policy

200 Model behaviour feedback, OpenAI https://openai.com/form/model-behavior-feedback

201 OpenAI - Bug Crowd, https://bugcrowd.com/openai

202 Richardson, Hansen, Google, Acting on our Commitement to safe and secure AI, 26 October 2023, https://blog.google/technology/safety-security/google-ai-security-expansion/; https://bughunters.google.com

203 Vela, Keller, Rinaldi, Google's reward criteria for reporting bugs in AI products, Google security, 26 October 2023, https://security.googleblog.com/2023/10/googles-reward-criteria-for-reporting.html

204 Vela, Keller, Rinaldi, Google's reward criteria for reporting bugs in AI products, Google security, 26 October 2023, https://security.googleblog.com/2023/10/googles-reward-criteria-for-reporting.html

ANNEX 4				
Company & System	**Reporting tools**		**Legal safe harbour**	
	Security	**Safety**	**Security**	**Safety**
Anthropic[205] (Claude 2)	Yes. Provides a reporting process.[206] Does not provide a bug bounty program with financial rewards.	Yes. Process to report model outputs that are inaccurate, biased, or harmful. [207]	Partial. For good faith research that complies with its responsible disclosure policy. Reserves discretion to decide if in good faith.	Unclear. Does not include an accompanying policy, legal safe harbour statement, or rewards.
Inflection[208] (Pi)	Partial. Operates a by-invitation closed bug bounty program for security and safety vulnerabilities. Stated it would create a publicly accessible bug bounty program in 2024.		Unclear.	
Mid-journey[209] (v6)	No.	Yes. Has a reporting tool for unsafe content.[210] Also has a specific process for copyright infringing material.[211]	Unclear. Does not include an accompanying policy or legal safe harbour statement.	

205 Acceptable Use Policy, Anthropic, https://www.anthropic.com/legal/aup; Bommasani, Scores for Anthropic on 2023 Foundation Model Transparency Index, https://github.com/stanford-crfm/fmti/blob/main/scoring/Anthropic%202023%20FMTI%20Scores.pdf, point 75.

206 Responsible Disclosure Policy, Anthropic, https://www.anthropic.com/responsible-disclosure-policy

207 Acceptable Use Policy, Anthropic, https://www.anthropic.com/legal/aup

208 Frontier Safety, Inflection, https://inflection.ai/frontier-safety

209 Terms of Service, Midjourney, https://docs.midjourney.com/docs/terms-of-service

210 Community Guidelines, Midjourney https://docs.midjourney.com/docs/community-guidelines

211 Terms of Service, Midjourney, https://docs.midjourney.com/docs/terms-of-service

ANNEX 4				
Company & System	**Reporting tools**		**Legal safe harbour**	
	Security	**Safety**	**Security**	**Safety**
Cohere[212] (Command)	Yes – combined for safety and security. No bug bounty program, but a publication of Cohere recommends bug bounties as an effective way to search and fix security weaknesses.[213]		Unclear. Usage policy allows stress testing of its API and adversarial attacks on the condition that violative generations are reported immediately.[214] No mention of legal safe harbour. Cohere publication also recommends safe harbours.	

212 Usage Guidelines, Cohere, https://docs.cohere.com/docs/usage-guidelines

213 The State of AI Security, Cohere, https://txt.cohere.com/the-state-of-ai-security/

214 Usage Guidelines, Cohere, https://docs.cohere.com/docs/usage-guidelines

Authors and Acknowledgments

This paper is written by Esme Harrington, Associate (AWO) and Mathias Vermeulen, Director (AWO).

We are grateful to reviewers Maximilian Gahntz, Claire Pershan, Nicholas Piachaud, Nick Botton, and the Social Science Research Centre's Data Fluencies Workshop 2024 organiser Dannah Dennis, chairs Inioluwa Deborah Raji and Amba Kak and all the attendees. We also want to thank the individuals who were interviewed over the course of this project.

This document was designed by Spitting Image, Bengaluru.