DEMOS

moz://a

OPEN HORIZONS

EXPLORING NUANCED TECHNICAL AND POLICY APPROACHES TO OPENNESS IN AI

ELIZABETH SEGER
BESSIE O'DELL

SEPTEMBER 2024

Published by Demos September 2024
© Demos. Some rights reserved.
15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

ABOUT THIS REPORT PAGE 4

ACKNOWLEDGEMENTS PAGE 4

EXECUTIVE SUMMARY PAGE 5

INTRODUCTION PAGE 7

SECTION 1: BACKGROUND PAGE 9

SECTION 2: WORKSHOP INSIGHTS PAGE 10

SECTION 3: OPEN QUESTIONS ABOUT OPENNESS PAGE 26

SECTION 4: MENU OF OPTIONS FOR GOVERNMENT FOR PROMOTING THE BENEFITS OF OPENNESS WHILE MITIGATING RISKS PAGE 31

CONCLUSION PAGE 40

ABOUT THIS REPORT

This paper is part of Demos' strategic focus area on ***'Trustworthy Technology'***. With emerging technologies transforming our world at an ever-faster pace, we work to build bridges between politicians, technical experts, and citizens to explore solutions, improve trust, and create policy to ensure our technologies benefit society.

This report is produced in partnership with Mozilla, and includes insights from a workshop held in June 2024 which convened global experts on open-source AI and AI 'openness'. It seeks to find places where we can push forward the policy discussion on safely pursuing openness benefits, even where there exists some disagreement in the broader debate. We do not pit risks against benefits, open against closed. Instead we look into what kinds of evidence should be collected in deciding when to release an AI model, how we might productively work to reduce any risks of openness where they do exist, and how we can use a variety of methods to pursue openness benefits. Finally, we translate these opportunities to a menu of policy options for responsibly harnessing the good of AI openness - to build and share AI technologies that are safe, trustworthy, and serve diverse needs.

ACKNOWLEDGEMENTS

We would like to thank the following people for offering their input, insight, and feedback at various stages throughout the drafting of this report:

Sumaya Akthar, Neel Alex, Dave Buckley, Joe Castellano, Peter Cihon, Harrison Durland, Gemma Edom, David Evan Harris, Shin-Shin Hua, Arjun Karanam, Sophia Knight, Sebastien Krier, Suryanash Mehta, Richard Moulange, Cailean Osborne, Udbhav Tiwari and Lucia Velasco.

Elizabeth Seger and Bessie O'Dell

September 2024

EXECUTIVE SUMMARY

Open-source software (OSS) development is a culture and set way of working that involves the free and open sharing of software projects for further study, modification, and use. For thirty years OSS has proliferated alongside (and often inside) private software, encouraging cooperation, fostering innovation and competition, nurturing talent, and improving software quality through community review. However, in the past couple of years, decades-old OSS tradition has been challenged by the development of increasingly capable AI. Concerns about the potential harms of malicious misuse has yielded heated debate about the prudence of open-sourcing AI with many arguing that some models pose too high of a risk to be made available for public download. The open-source AI debate has been a high-stakes dialogue, measuring up harms against benefits and risks against worries.

Yet despite what, at its start, often felt like a stark open-v.-closed, us-v.-them standoff, the open-source AI debate has evolved productively into collaborative conversations. This has included noted progress toward establishing a clear definition for open-source AI, broader consideration of a spectrum of model sharing strategies, and exploration of numerous dimensions of AI openness expanding beyond core considerations of model access.

Building on insights from the ''Open Horizons'' workshop hosted by Demos and Mozilla on June 6th 2024 – a follow-up to Demos's October 2024 open-source policy workshop and report[1] – this report reflects on areas of emerging consensus, persisting disagreement, and unanswered questions to start thinking about next steps for safely pursuing openness benefits. This report does not pit risks against benefits. Instead we look into what kinds of evidence should be collected in deciding when to release an AI model, and how we might productively work to reduce any risks of openness where they do exist.

THIS REPORT YIELDS THE FOLLOWING KEY INSIGHTS:

- Risk mitigation strategies and guardrails need to be implemented throughout the AI value chain and through improvements to systemic AI safety, not just at the point of model release.
- Urgent work is needed to develop clear threat model and model evaluation standards.
- We can seek alternative methods for pursuing AI openness goals instead of, or alongside, model sharing for when model access restrictions are in place.
- There are promising technical solutions to explore which can help mitigate some of the risks of openness.

[1] Ball, J., & Miller, C. (2024). Open Sourcing the AI Revolution: Framing the debate on open source, artificial intelligence and regulation. Demos. Retrieved August 18, 2024, from https://demos.co.uk/research/open-sourcing-the-ai-revolution-framing-the-debate-on-open-source-artificial-intelligence-and-regulation/

- The impacts of restricting access to highly capable AI models on competition and market concentration need to be researched and clarified.
- Unclear liability rules and safety standards stifle open innovation and can yield less safe technologies.

Based on these key insights we develop a non-exhaustive menu of policy options for governments. These options outline forward-looking areas of (1) investment and (2) potential regulatory interventions, by which the government might promote the benefits of openness that facilitate innovation, enable digital autonomy, fuel competition, and attract talent while mitigating potential harms. We explore the strengths and weaknesses of these options in the main text.

(1) INVESTMENT OPTIONS:

- Provide financial support for open-source projects and ecosystems. This includes providing support for open-source safety testing ecosystems for smaller developers who may not have access to tools, standards, or necessary compute resource.
- Invest in digital public infrastructure such as open national AI models, public compute infrastructure, and open data libraries.
- Invest in clear threat modelling exercises involving domain experts to underpin targeted risk mitigation policy.
- Investigate economic impacts of open-sourcing AI models and of restricting model access.
- Investigate guardrails that can be deployed throughout the AI value chain to reduce risks of AI openness.
- Invest in incentive structures such as large rewards programs to promote AI safety and social benefit breakthroughs.

(2) REGULATION OPTIONS:

- Set public sector procurement standards for transparency to incentivise greater AI openness.
- Introduce exemptions from AI regulation for models that meet a certain standard of openness/transparency in order to incentivise greater transparency and information sharing.
- Clarify liability legislation and establish openness as a condition for transferring liability downstream.
- Define part of AI Safety Institute (AISI)'s role as providing safety evaluation support for startup and open-source ecosystems - e.g. by providing resources, tools, and safety evaluation standards.
- Establish / reform government open data policy.
- Incorporate democratic processes into government decision-making around AI.

INTRODUCTION

On June 6th, 2024 Demos hosted a workshop in partnership with Mozilla to investigate nuanced policy and technical approaches for pursuing the benefits of openness in AI. The "Open Horizons" workshop brought together a diverse range of experts and was structured to build upon an incredibly rich past couple of years of discussion pertaining to open-source AI.

Generally speaking, to 'open-source' an AI model is to make the model freely and publicly accessible for anyone to study, modify, use, and share. The practices, norms, and values surrounding open-source AI stem from decades of fruitful open-source software tradition (the free and open sharing of software projects for further study, modification, and use) and academic norms of open access research publication. For thirty years OSS has proliferated alongside and often inside private software, encouraging cooperation, fostering innovation and competition, nurturing talent, and improving software quality.[2]

However, in recent years, the OSS tradition has been challenged by the development of increasingly capable AI. Concerns about the potential harms resulting from exposure of vulnerabilities and malicious misuse of these models has yielded heated debate about the prudence of open-sourcing AI, with some stakeholders arguing that certain models may pose too high of a risk to be made available for public download.[3] The open-source AI debate has therefore been a high-stakes dialogue measuring up harms against benefits, and risks against worries.[4]

Yet recently, the discussion on open-source AI has evolved from what, at its start, often felt like a deeply polarised open-v.-closed, us-v.-them standoff, to a fruitful and collaborative dialogue. We have seen noted progress towards a clear definition of 'open-source AI' spearheaded by the Open Source Initiative (OSI)[5] and broader consideration of a spectrum of model-sharing methodologies, from fully open to fully closed.[6] The concept of AI openness has also expanded beyond questions of model access and use to consider access to data, compute resources, and community support.[7] Works such as the Carnegie Endowment's consensus investigation,[8] Demos's policy framing report,[9] and the NTIA's recent multistakeholder review[10] have helped

[2] Engler, A. (2021). How Open-Source Software Shapes AI Policy. AI Governance Report, Brookings. Retrieved August 10, 2024 from https://www.brookings.edu/articles/how-open-source-software-shapes-ai-policy/

[3] Seger, E. et al. (2023). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Retrieved July 20, 2024, from https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[4] See section 2.2 on for an overview of open-source risks and benefits.

[5] OSI (2024). The Open Source AI Definition - Draft 0.0.8. Retrieved 20 July, 2024, from https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8

[6] Solaiman, I. (2023). The Gradient of Generative AI Release: Methods and Considerations. Retrieved 20 July, 2024, from https://doi.org/10.1145/3593013.3593981

[7] Marda, N. (2024). The Columbia Convening on Openness and AI: Technical Readout. Retrieved July 18, 2024, from https://foundation.mozilla.org/en/research/library/technical-readout-columbia-convening-on-openness-and-ai/

[8] Bateman, J. et al. (2024). Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance. The Carnegie Endowment for International Peace. Retrieved 24 July, 2024, from https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en

[9] Ball, J., & Miller, C. (2024). Open Sourcing the AI Revolution: Framing the debate on open source, artificial intelligence and regulation. Demos. Retrieved August 18, 2024, from https://demos.co.uk/research/open-sourcing-the-ai-revolution-framing-the-debate-on-open-source-artificial-intelligence-and-regulation/

[10] NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

shift a polarised discourse into a collaborative effort to seek common grounds of agreement from which to progress. The Centre for the Governance of AI's report on open-sourcing highly capable foundation models encouraged broader thinking about the variety of mechanisms that yield openness benefits when model access restrictions are in place.[11] Additionally, the Partnership on AI, in collaboration with GitHub, have produced a useful taxonomy of the various actors along the AI value chain[12] (not just foundation model developers) who have levers they can pull to influence the risks and benefits of AI openness.[13]

WHAT DOES THIS REPORT DO?

Building on insights from the June 6th Demos/Mozilla "Open Horizons" workshop, this report follows the prevailing theme of collaborative and productive discussion. The report presents a broad understanding of openness that extends beyond model access considerations, and that appreciates the importance of maintaining openness around AI while also acknowledging associated risks.

The report's unique contribution is to find places where we can push forward the policy discussion for safely pursuing openness benefits even where some disagreement in the broader debate still exists. To do so, we do not pit risks against benefits or try to define when, precisely, a model should be released. Instead we look into (a) what kinds of evidence should be collected in deciding when to release a model, and we ask: (b) how we might productively work to reduce any risks of openness where they do exist and (c) where access restrictions are put in place, how might we pursue the benefits of openness by alternative means.

In this report we identify different actors with levers to pull, and we take the next step in translating opportunities to policy options. The report's main contributions are a list of open questions about openness (section 3) and a menu of policy options (section 4) for how the state can support and catalyse the pursuit of the benefits of AI openness for businesses and citizens while also being mindful of the risks.

[11] Seger, E. et al. (2023). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Retrieved July 20, 2024, from https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[12] The 'AI value chain' can be defined e.g., as 'the organisational process through which an individual AI system is developed and then put into use (or deployed)'. See: Engler, A. C. and Renda, A. (2022). Reconciling the AI Value Chain with the EU's Artificial Intelligence Act. Retrieved August 21, 2024, from https://openfuture.eu/wp-content/uploads/2024/03/220930CEPS-In-depth-analysis-AI-act-value-chain.pdf

[13] Srikuman, M., Chang, J. & Chmielinski, K. (2024). Risk Mitigation Strategies for the Open Foundation Model Value Chain. Retrieved July 20, 2024, from https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

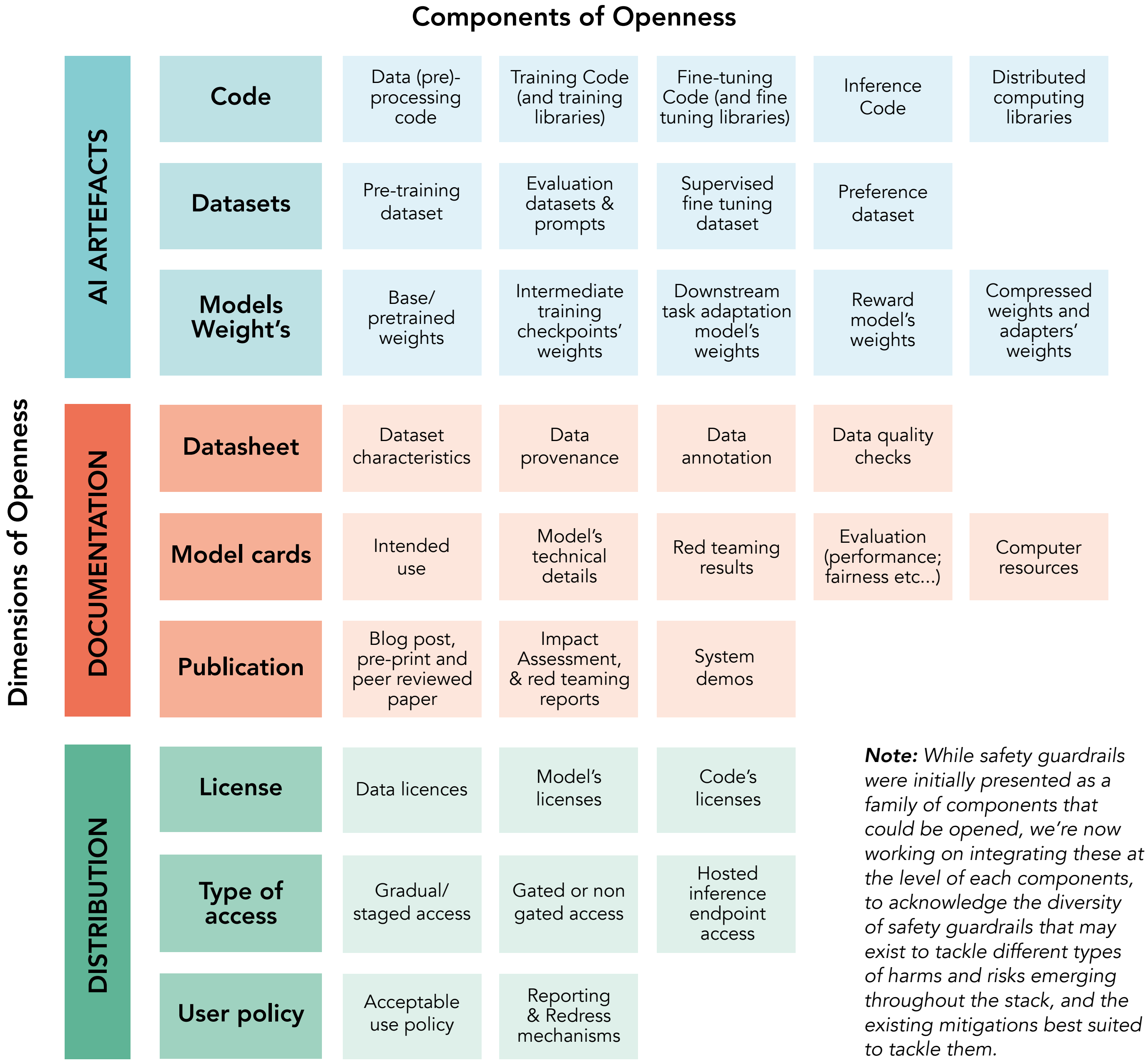
SECTION 1
BACKGROUND

1.1 WHAT IS AI OPENNESS?

For this workshop and report we build on a concept of AI openness initially explored during the 2024 convening on AI Openness co-hosted by the Mozilla and the Columbia Institute of Global Politics (hereafter 'the Colombia Convening').

AI openness is still in want of a settled definition. However as presented at the Colombia Convening, **AI openness can be generally understood as the broad public availability of key artefacts and documentation from AI across the AI stack**. Figure 1, reproduced from the Columbia Convening technical readout, presents an initial effort to taxonomise the various dimensions along which AI openness might be evaluated. It includes model release and distribution options, licensing, access to AI artefacts and documentation, and transparency into safety guardrails.

FIGURE 1
OPENNESS ACROSS THE AI STACK



AI ARTEFACTS	Code	Data (pre)-processing code	Training Code (and training libraries)	Fine-tuning Code (and fine tuning libraries)	Inference Code	Distributed computing libraries
	Datasets	Pre-training dataset	Evaluation datasets & prompts	Supervised fine tuning dataset	Preference dataset	
	Models Weight's	Base/ pretrained weights	Intermediate training checkpoints' weights	Downstream task adaptation model's weights	Reward model's weights	Compressed weights and adapters' weights
DOCUMENTATION	Datasheet	Dataset characteristics	Data provenance	Data annotation	Data quality checks	
	Model cards	Intended use	Model's technical details	Red teaming results	Evaluation (performance; fairness etc...)	Computer resources
	Publication	Blog post, pre-print and peer reviewed paper	Impact Assessment, & red teaming reports	System demos		
DISTRIBUTION	License	Data licences	Model's licenses	Code's licenses		
	Type of access	Gradual/ staged access	Gated or non gated access	Hosted inference endpoint access		
	User policy	Acceptable use policy	Reporting & Redress mechanisms			

Note: *While safety guardrails were initially presented as a family of components that could be opened, we're now working on integrating these at the level of each components, to acknowledge the diversity of safety guardrails that may exist to tackle different types of harms and risks emerging throughout the stack, and the existing mitigations best suited to tackle them.*

Reproduced from the Columbia Convening Technical Readout

The taxonomy taken from the Columbia Convening is a work in progress. In this report we propose **Figure 1** can be expanded further to incorporate accessibility of compute resources, accessibility and support for participating in development communities and collaborative initiatives, and the transparency of and mechanisms for inputting to AI governance processes.

"AI openness" is distinct from "open-source AI".[14] A specific definition of Open-Source AI is being developed in a co-design process coordinated by the Open Source Initiative (OSI). The current definition, which OSI aims to settle by October 2024, refers to the availability of a model

[14] For a thorough discussion of the difference between often confused terms such as open access, open-source, open science, open licence, open knowledge, and open collaboration in relation to AI see White et al. (2024). The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. Retrieved from https://arxiv.org/abs/2403.13784

for public download under open-source licence such that anyone is able to freely use, study, modify, and share the model.[15] Open-source is, however, only one kind of model distribution option, and model distribution is one component of AI openness.

In this paper, when we refer to "open-source AI" we specifically refer to the draft definition set out by the OSI including downloadable models and licence requirements. We refer to "downloadable models" when discussing downloadable model access with licences aside, and "AI Openness" to refer to the wider ecosystem of openness that can exist around AI development, distribution, and governance.

The advantage of having a broad concept of AI openness is that it allows us to delineate between different notions of openness that we often assume are linked such as "openness as transparency", "openness as access" and "openness as permission" and to explore how different components of openness can be prioritised by different stakeholders. Our preliminary discussion on the topic is outlined in **Table 1**.

TABLE 1
COMPONENTS OF OPENNESS TYPICALLY PRIORITISED BY DIFFERENT STAKEHOLDERS

STAKEHOLDER	PRIORITISED COMPONENTS OF AI OPENNESS
Downstream Developers	• Open-source licences. • Downloadable model access. • Documentation / research publications.
Researchers / Academia	• Downloadable model access. • Documentation / research publications. • Dataset access. • User logs. • API mode access.[16]
Consumers	• API model access. • Point of use guardrails.
Creative Industry	• Data set transparency - for copyright holders (publishers, journalists, artists, etc…) to identify what their work is being trained on.
Civil Society	• Transparency into audit guardrails. • Datasets and Datasheets transparency - to address concerns of bias and discrimination).

[15] For specific conditions and model component access requirements see: OSI (2024). The Open Source AI Definition - Draft 0.0.8. Retrieved 20 July, 2024, from https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8

[16] For some, but not all, research cases, free and unlimited API model access may be more valuable than downloadable access. For example, when doing large-scale adversarial testing or testing it can be prohibitively expensive for researchers to repeatedly run large models downloaded to their own systems. It is more feasible for researchers to study shared neural networks with shared compute resource. E.g. see https://www.khoury.northeastern.edu/research_projects/national-deep-inference-fabric-ndif/

1.2 A BRIEF OVERVIEW OF OPENNESS BENEFITS AND RISKS

There are various benefits and risks to making model models available for public download that have formed the basis of the AI open-source debates. The following sources provide a extensive discussion of the risks and benefits:

- **Seger et al. (2023).** Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives.[17]
- **Bommasani et al. (2023).** Considerations for Governing Open Foundation Models.[18]
- **Kapoor et al. (2024).** On the Societal Impact of Open Foundation Models.[19]
- **Basdevant et al. (2024).** Towards a Framework for Openness in Foundation Models: Proceedings from the Columbia Convening on Openness and Artificial Intelligence.[20]
- **NTIA Report (July 2024).** Dual-Use Foundation Models with Widely Available Model Weights.[21]

For ease of reference, **Table 2** provides a high level summary of key openness benefits and risks.

TABLE 2
HIGH-LEVEL SUMMARY OF OPENNESS BENEFITS AND RISKS

BENEFITS	• Improving AI safety and security by allowing a wider and more diverse community of developers and researchers to appraise models, identify and work to remedy vulnerabilities. • Accelerating beneficial AI progress and innovation (e.g. in algorithmic efficiency) by democratising AI development and creating a collaborative environment of AI developers and makers. • Enabling digital autonomy by making advanced technologies more accessible to actors that might otherwise lack resources to develop them independently. This reduces reliance on proprietary systems controlled by a few large tech companies or powerful nations. • Improving competition, reducing market concentration, and mitigating vendor lock in - *(some aspects of this point are disputed - see below).*
RISKS	• Bypassing safeguards against misuse. • Misuse without oversight / ability to monitor model use. • Dissemination of dangerous capabilities. • Possible introduction of new dangerous capabilities by fine-tuning models. • Perpetuation of safety vulnerabilities and model flaws. • No take-backs - once a model is made publicly downloadable there is no way to rollback downloaded copies.

[17] https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf
[18] https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf
[19] https://arxiv.org/abs/2403.07918
[20] https://arxiv.org/pdf/2405.15802
[21] https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

The Demos/Mozilla Open Horizons workshop did not dive into the intricacies of these benefits and risks, though participants offered several insightful observations about the benefits and risks of AI openness.

1. Risks are primarily posed by the sharing of model components (predominantly, model weights), while the benefits increase with greater openness across all dimensions of AI openness.

The asymmetry arises because the risks are concentrated in the immediate capabilities of the model (encapsulated in weights) and how those capabilities can be misused. Meanwhile, the benefits often come from the broader ecosystem of knowledge, tools, and practices that openness fosters.

2. The "No Take Backs" point about publicly downloadable model access is often overstated.

A commonly raised reason to be cautious about open-sourcing is that once a model is made publicly downloadable, there is no way to reverse the decisions if harms begin to manifest. The model is out there in the world. It is true that once an AI model is openly released for download, a wholesale rollback of all existing model copies is impossible. However, hosting platforms can make it very difficult for malicious actors to find downloadable copies of a model by removing it from their platforms. This difficulty in downloading copies of a model would probably pose a sufficient barrier to the majority of malicious actors looking for an easy way to cause harm. That said, it is not the common malcontent we are worried about when considering the more extreme risks that could be posed by highly-capable systems in the future. The greater concern pertains to the malicious intentions of well-resourced and strongly motivated actors like politically motivated states, who may not be so easily thwarted in obtaining model access. The question then becomes: are such well-resourced actors given a significant uplift by the availability of open-source tools, or would they likely be able to develop such tools themselves?

3. There is disagreement about the economic impacts of open-sourcing AI models on market concentration, competition, and vendor lock stemming from a lack of information and clarity.

There seem to be mechanisms by which open-sourcing can convey both positive and negative impacts on competition, market concentration, and control. For example, in the near term, open-sourcing allows downstream developers to build on and innovate using highly capable models they would not have been able to afford to train themselves. This is good for competition at the application development level. But open-sourcing highly capable models might also provide large developers with a mechanism for further strengthening their positions as industry leaders. In the longer term, companies that open-source large foundation models establish their tools and architectures as standard in the ecosystem. The downstream innovations that build on open-sourced models can be readily incorporated back into the original developer's products and platforms. Additionally, open-source communities serve as valuable talent pools, providing companies with potential hires who are already well-versed in their tools and models. These conditions do not necessarily constitute a breach of antitrust law - overall they might enhance efficiency and product quality for consumers - but the point stands that open-sourcing large, expensive-to-train foundation models is not an entirely selfless act.

Furthermore, whether open-sourcing reduces the cost to customers for switching between platforms ('switching costs') depends on the availability of suitable alternatives and the interoperability of existing tools. Open-source facilitates interoperability by providing transparent standards, customizable code, and fostering a community-driven approach to solving integration challenges. However, interoperability might also be achieved for proprietary models by legal requirement, for example through antitrust or consumer protection law. However, implementing such legal requirements is a considerable challenge and requires evidencing harm to competition and consumers.

Finally, while it is important that open-source alternatives exist, many business consumers may not be motivated to switch to open-source models irrespective of their availability. Large providers have the resources and internal expertise to develop and test model applications, whereas it would be a considerable cost for many business consumers to establish comparable expertise internally. These providers can also offer support level agreements (SLA's) - if something goes wrong the provider supplies support and maintenance services to the customer. Demand for proprietary models will therefore likely remain high as the more economical decision for many consumers. These benefits of using proprietary models will, of course, only be enjoyed by those who can afford it. The availability of open-source models is important for democratising access to AI tools for all, and there are businesses such as RedHat that focus on providing customer support and maintenance for open-source software.

Participants agreed that this disagreement about the economic and market impacts of open-sourcing AI stems from a lack of information and specificity about what technologies and key players we have in mind. More research is needed to clarify the technical and economic dynamics at play.

1.3 AREAS OF HIGH-LEVEL CONSENSUS THAT PROVIDE A PRODUCTIVE STARTING POINT FOR DISCUSSION

At the start of the workshop we identified high-levels of consensus among the diverse group of attendees to help ground the conversation and to provide a common point of departure. The points of high-level consensus were as follows:

1. "Open-source AI" means something very specific referring to how a model is released and licensed for downstream development and use.
2. There is a spectrum of model release options from fully–closed to fully-open that involve tradeoffs between risks and benefits.
3. AI openness is an important concept, distinct from open-source, that is about more than model access. It is about openness through the AI stack (including data, compute, documentation, and licensing) and extends to discussions about transparent decision making, accessible education, diverse community participation, and how (and by whom) AI and institutions are governed.
4. AI is a very broad category of technologies spanning a wide range of capabilities, complexities, model sizes, and application spaces.
5. The risks and benefits of openness will vary for different models and applications.
6. It is conceivable that one day we develop a model that we would not want to be open-sourced or made available for public download due an unacceptable level of risks posed by the model.[22]

We expect that this is not an exhaustive list of the consensus points that existed within our group. Further investigation surely would have yielded further points of common ground with a higher resolution analysis of where and why our alignments start to fray.

Such an investigation was conducted by the Carnegie Endowment for International Peace in April 2024 and recently published.[23] While the Carnegie Endowment report goes into greater detail, the areas of agreement we identified strongly align.

[22] This statement purposefully does not attempt to articulate specific thresholds, capabilities, or expected timelines. The timeline on "one day" is intentionally left vague as there was disagreement in the group about how close we are to developing models that surpass an unacceptable risk threshold or, indeed, as to whether some such models already pass the threshold. The key point here is that no one denies the possibility of producing a model too dangerous to share. Our group's consensus is also consistent in the U.S. NTIA report conclusion that "current evidence is not sufficient to definitively determine either that restrictions on such openweight models are warranted, or that restrictions will never be appropriate in the future." NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[23] Bateman, J. et al. (2024). Beyond Open vs. Closed: Emerging Consensus and Key Questions for Foundation AI Model Governance. The Carnegie Endowment for International Peace. Retrieved 24 July, 2024, from https://carnegieendowment.org/research/2024/07/beyond-open-vs-closed-emerging-consensus-and-key-questions-for-foundation-ai-model-governance?lang=en

SECTION 2
WORKSHOP INSIGHTS

The Open Horizons workshop primarily aimed to move the conversation beyond the open v. closed and risks v. benefits by identifying opportunities for realising openness benefits even where disagreements may still exist. We pursued three lines of discussion pertaining to (2.1) how model release standards should be informed, (2.2) how we might productively work to reduce risks from openness where they do exist, and (2.3) where access restrictions are put in place, how might we pursue the benefits of openness by alternative means.

2.1 INSIGHTS ON SETTING STANDARDS FOR MODEL RELEASE

Our goal here was not to define standards or thresholds for model release but to ask the prerequisite questions about first steps that ought to be taken. This included questioning, for example: what kind of evidence should we be looking for, how should risks and benefits be evaluated, and what limitations might we encounter? Our discussion yielded the following points:

2.1.1 More work is needed to develop and standardise benchmarking and benchmark validation tools to evaluate model capabilities.

Current AI policy often stipulates compute thresholds to trigger more stringent model evaluation and safety requirements. While this approach provides a valuable starting point for identifying potentially high-risk models, it offers only a rough and increasingly tenuous proxy for capability.[24,25] Recent advances in fine-tuning techniques, such as Low Rank Adaptation (LoRA), have demonstrated that the performance of smaller models can be significantly improved by optimising model weights using specialised data, including outputs from more capable models.[26] Consequently, model capability, rather than size, should be the primary metric for evaluating potential harmful misuse.

[24] Heim (2024). (Training) Compute Thresholds: Features and Functions in AI Governance. Retrieved July 20, 2024, from https://arxiv.org/pdf/2405.10799

[25] Cohere for AI (2024). Exploring the Role of Compute-Based Thresholds for Governing the Risks of AI Models. Retrieved August 12, 2023 from https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf

[26] T. Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs. Retrieved July 20, 2024, from https://arxiv.org/abs/2305.14314

However, as noted by the Department for Science, Innovation & Technology (DSIT), there is no widely-accepted definition of 'capability' in the field of AI, despite general-purpose AI systems often being described in terms of their quantitative and qualitative capabilities.[27] This challenge is compounded by the multitude of model benchmarks used to measure performance[28] and the difficulty in extrapolating overall capability and future performance from specific task results.[29] To address this, work must continue to improve and standardise benchmarks for consistent evaluations between models.[30] A greater challenge yet lies in validating these benchmarks themselves, which involves verifying the accuracy and reliability of the AI benchmarks used to assess model capabilities.

Greater openness about benchmarks - i.e. including information sharing between organisations and the maintenance of open-source benchmarks - can facilitate alignment and consistent implementation and identify gaps in current assessment methods.[31] The potential downsides of greater openness about benchmarks and evaluations is that developers can design to the test - and thereby, models are overfitted to score well on benchmarks without those high scores reflecting model capability or safety in broader contexts. Some benchmarks could also be used to infer information about model development that a country may not wish to be public, e.g. for models being developed for application in a national security setting. There is therefore a question of what information about benchmark data and evals should be available to everyone (including developers) and what should only be known by external auditors or evaluators.

It is a core recommendation of the U.S. NTIA Report on open weight foundation models for the US government to build internal capacity and invest in developing benchmarks and definitions for monitoring model development and to guide policy action.[32] The US and UK AISIs might collaborate on this front.

2.1.2 We need to develop clear threat models to guide model release decisions.

Stipulating clear threat models to guide AI model evaluation and release strategy is crucial for effective risk management. By defining specific pathways to harm, organisations can focus their efforts on the most critical threats, prioritise resources efficiently, and develop tailored mitigation strategies. This targeted approach makes the risk assessment process more manageable dealing with complex AI systems that could potentially pose a wide range of risks, some of which may be missed in initial model evaluations. The NTIA report similarly recommends that the U.S. federal government develop and maintain a set of risk portfolios, indicators, and thresholds.[33]

Clear threat models provide a common framework for discussing potential dangers and aligning various parties on the specific risks being addressed. They can also help to guide the development of more relevant and targeted benchmarks for evaluating model safety and performance in critical areas. Threat models also help us identify points throughout the pathway, aside from model development and release, where interventions can help mitigate threats.

During the workshop, it was discussed that in practice, threat model development could start with the publication of responsible scaling policies (RSPs). This could be followed with later reverse-engineering the threat models by identifying threat actors and outlining pathways to harm. During

[27] UK Department of Science, Innovation, and Technology (May 2024). International scientific report on the safety of advanced AI: interim report. Retrieved July 12, 2024, from https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai/international-scientific-report-on-the-safety-of-advanced-ai-interim-report

[28] For example, the Beyond the Imitation Game benchmark (BIGbench) currently consists of 214 tasks, which were initially contributed to by 450 authors across 132 institutions, and which includes task topics which draw upon problems from linguistics, child development, maths, physics, and more. See: Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Retrieved July 20, 2024, from https://arxiv.org/abs/2206.04615

[29] UK Department of Science, Innovation, and Technology (May 2024). International scientific report on the safety of advanced AI: interim report. Retrieved July 12, 2024, from https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai/international-scientific-report-on-the-safety-of-advanced-ai-interim-report

[30] For a detailed and accessible overview on benchmarking AI, see Janapa Reddi (2024). 'Benchmarking AI' in Machine Learning Systems with TinyML.

[31] Ibid.

[32] NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[33] Ibid.

the process of developing clear threat models it is important to seek input from stakeholders with specific subject area expertise (e.g. virologists on biological risks, and AI ethicists and civil society on bias and discrimination) to ensure key factors and mechanisms are not overlooked. Part of this process will involve clarifying what constitutes harm, and to whom.

Finally, specifically with respect to modelling threats resulting from AI openness, we need to establish a more fine-grained understanding of what capabilities are conveyed to malicious actors through the release of different combinations of model components. This is to ensure that any model access restrictions put in place as a result are not unnecessarily restrictive by more precisely targeting the source of risk.

2.1.3 Risks should be evaluated as 'marginal uplift' to malicious actors.

'Marginal uplift' or 'marginal risk' describes the additional risk the technology poses through intentional misuse beyond that posed by pre-existing technologies or closed-source versions.[34] It is about what additional capability an AI model conveys to a malicious actor above what they could achieve with existing technologies. For example, we might ask: how much more of a challenge to information environment security do AI image generators pose above and beyond that already posed by Photoshop? And do large language models give better instructions on how to build a bomb with garden fertiliser than you could find with an internet web search? Attending to marginal risk is important to prevent fear mongering and to ensure recommended interventions are proportional to the threat posed.

Workshop participants generally agreed that risks of open-source should be appraised in terms of marginal risk. The U.S. NTIA report on dual use foundation models noted similar agreement among their respondents,[35] though some of the Demos/Mozilla workshop participants cautioned that we still need to clearly define a stable acceptable risk threshold against which marginal risk is appraised in order to avoid a "boiling frog" scenario. That is, if incremental improvement in model capability adds only a minor addition of marginal risk compared to the last version (the last version being a pre-existing technology), then we may find that we've layered marginal risk upon marginal risk until, before we realise it, we are publicly releasing really quite dangerous technologies.

2.1.4 Should we measure benefits too?

One participant suggested that if we are measuring risk, we should also be trying to measure benefits - identifying clear benefit pathways and identifying what aspects of openness are most important to specific stakeholders; only then can we really measure up risks against benefits. There was some pushback in the group that the onus should not be on having to prove benefits - instead we should focus on restrictions based on the severity of harm, and the difficulty of mitigating the harm. The counter response was that if clear evidence of increased marginal risk is needed to justify access restrictions, then why should not clear evidence of marginal benefits be required to defend a default open position?[36] When thinking from the perspective of enforcing regulation (e.g. antitrust law to preserve competition and protect consumers) both the benefits and risks of the proposed interventions (its implementation and non implementation) must be analysed and compared.

2.1.5 Don't let model release standards distract from other risk mitigation strategies.

Finally, while it is important that we work to define clear model release standards based on specific threat models and well-established and verifiable capability benchmarks, we should not over-index on controlling model release as a mechanism for mitigating risks from highly capable AI models.

[34] Kapoor, S. et al. (2024). On the Societal Impact of Open Foundation Models, from https://arxiv.org/pdf/2403.07918v1

[35] NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[36] The U.S. NTIA report (ibid.) has also noted that there is significant uncertainty around both the future harms and benefits of any AI application, so indicators for both should be taken into account for risk benefit calculations.

Our methods of model evaluation are imperfect, and even where more restrictive model sharing practice is enforced, occasional model component leaks should be anticipated. Therefore, while we should consider building good model release standards, we must also think about guardrails that can be put in place throughout the AI stack and about harm mitigation at point of use and at other points in a given threat model. For example, with respect to mitigating AI biorisk (e.g. using large language models to generate DNA sequences for virulent pathogens) one risk mitigation strategy is to try to prevent the distribution of capable AI models. Another, and more likely to be effective risk mitigation measure, is to track and restrict the distribution of expensive specialist DNA synthesis machinery.

Finally we should investigate where else in the AI stack guardrails might be implemented to mitigate risks from AI openness. We attend to this discussion in the following section.

2.2 INSIGHTS ON MITIGATING RISKS FROM OPENNESS

Concerns about openness stem from the risk that making a model publicly downloadable exacerbates misuse and vulnerability risks by allowing malicious actors to bypass safeguards or modify models to put them to nefarious ends. By investigating how we can reduce the risks of downloadable model access, we make the risk-benefit trade-off less steep.

The aim here is to build in guardrails against misuse and vulnerability proliferation across the AI stack, not just at the point of model release.

Workshops participants commented on the following solution spaces that we encourage AI actors and governments to investigate further:

2.2.1 Potential technical solutions for mitigating risks

Research into technical solutions for reducing risks from openness aim at either making it more difficult to misuse models or making it possible to "take-back" open-sourced models if things go awry. Such technical solutions tend to be conceptual or in development (as opposed to being readily available) and would benefit from additional research investment. Some promising examples include:

Retrieval Models: It is technically possible to equip AI models with the ability to directly access (e.g., retrieve) a large database as they perform predictions[37] - a functionality initially introduced to maintain model performance, whilst improving model efficiency and decreasing computational demand. So-called 'retrieval-based deep learning' offers a possible technical solution to AI model governance that bridges the gap between open-sourced and closed-sourced models. In effect, this approach allows for the partitioning of ''safe'' versus ''unsafe'' model capabilities into (literally) different sections of the model.[38] In practice, this could allow a model developer to open-source parts of the model that are not likely to be dangerous, and to exclude the model from being able to access any knowledge (contained within the database the model is retrieving from) that could lead to potentially dangerous capabilities. As an example, if a model had a 'biology' section, which might be capable of allowing for the creation of biological weapons, then a company could exclude its knowledge of biology from the database from which the model retrieves its knowledge and open-source the rest of the model.[39]

There are some limitations to retrieval-based deep learning such as contextual rigidity (retrieval

[37] For example, Google Deepmind introduced the Retrieval-Enhanced Transformer (RETRO) model in 2021. See: Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022, June). Improving language models by retrieving from trillions of tokens. In International conference on machine learning (pp. 2206-2240). PMLR.

[38] OpenMined (2024). Response to the National Telecommunications and Information Administration (NTIA) Request For Comment (RFC) on Dual Use Foundation Artificial Intelligence (AI) Models with Widely Available Model Weights. Retrieved July 12, 2024, from https://www.regulations.gov/comment/NTIA-2023-0009-0334

[39] Ibid.

models often have narrower domains of application, though some research indicates the performance gap might be overcome[40]) and the need for access to the source data. However, perhaps the greatest challenge to overcome is the lack of infrastructure available to scale retrieval-based models. Open-source models, for example, are not typically trained to do retrieval augmented generation, and open-source frameworks like PyTorch and Tensorflow do not ship with the ability to learn retrieval augmented generation. As a result there is considerable social momentum within AI model development communities pushing innovation toward non-retrieval-based strategies. This momentum might be overcome by further funding research and open-source development projects in this area.

Self-destructing models: The goal of self-destructing models is to make models extremely difficult to tamper with. These models are optimised to work well for a constrained task, but if an adversary tries to repurpose the model by fine-tuning, performance tanks to the same level as an untrained model parameters, essentially forcing the actor to start from scratch. In this way models may be made fully downloadable while significantly reducing the cost to malicious actors in repurposing the model for unintended purposes. The downside is that self-destructing models substantially decrease the benefit of openness that allows downstream developers to innovate and iterate on trained models through fine-tuning. Self-destructing models do, however, still offer transparency to facilitate research and external evaluation.

"Baked in" watermarks: Watermarks are used to identify artificially generated content. Depending on how watermarks are built into a model they can be more or less easy to remove or bypass given access to the model.[41] If, for example, the watermark is added post hoc in the inference code (the code that tells the model to run) then the watermark could easily be removed by deleting the line of code. There are other methods that may be able to more effectively "bake in" the watermark. For example, if a model were exclusively trained on watermarked images, then the only way to remove the watermark for generated images would be to completely retrain the model.[42] However, due to restricted data set size using this method, other issues with data diversity and bias are likely to arise. Another possibility is to merge watermarking into the image generation process itself through adjustments to the pretrained model, though details of this process need to be kept secret to prevent removal.[43]

The possibility of implementing irremovable watermarks for artificially generated content is a promising area of research but much more work is needed. Watermarking text is a particularly difficult challenge.

Formal verification methods: Formal verification would use mathematical and logical methods to prove that an AI system behaves according to specific safety and performance criteria. Robust formal verification would help reduce the risk of open-sourcing models by proving the models will not engage in harmful behaviours. However, while formal verification has been successfully applied in some areas of computer science and software development, application to large and complex foundation models (like large language models) is currently infeasible. Application to smaller and narrow predictive machine learning systems (expert systems) is most promising.[44] Significant research and development is needed.

Secure enclaves: A secure enclave is a type of computer chip, manufactured with a "private key" burned into the chip in a way that no-one else can know its value. This allows an external party to encrypt information using a "public key" with the knowledge that the only computer in the entire

[40] Borgeaud. S. et al. (2021). Improving language models by retrieving from trillions of tokens. Retrieved August 14, 2024, from https://arxiv.org/abs/2112.04426

[41] Srinivasen, S. (2024). Detecting AI fingerprints: A guide to watermarking and beyond. Brookings. Retrieved July 20, 2024 from https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/

[42] Yu, N., Skripniuk, V., Abdelnabi, S., & Fritz, M. (2021). Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data. Retrieved July 20, 2024, from https://ieeexplore.ieee.org/document/9711167

[43] Fernandez, P. et al. (2023). The Stable Signature: Rooting Watermarks in Latent Diffusion Models. Retrieved July 20, 2024, from https://arxiv.org/pdf/2303.15435

[44] Kohli, P., Dvijotham, K., Uesato, J. & Gowal, S. (2019). Identifying and eliminating bugs in learned predictive models. Retrieved July 12, 2024, from https://deepmind.google/discover/blog/identifying-and-eliminating-bugs-in-learned-predictive-models/

world that could decrypt this information and use it is the secure enclave. The application of secure enclaves can facilitate structured access to closed models, enabling some benefits traditionally associated with open models (dataset/model/API transparency) to be realised whilst maintaining protection of IP.

Loading a model into a secure enclave also opens up the possibility for more flexible governance frameworks. In the simplest case, a third-party auditor or regulator could send audit code/data to execute against the model inside the enclave. The third-party would see the results of the audit, but not the underlying model training data/weights/APIs. In principle, this mechanism could be extended to include many auditors which, when coupled with an appropriate governance framework, could allow closed models to reap the external transparency benefits currently only available to open models.[45]

GPU enclaves are currently in tech preview[46] and their application to AI governance use cases is being actively piloted.[47]

2.2.2. Mitigating risk throughout the AI value chain

Participants discussed a variety of levers that players throughout the AI value chain could pull to mitigate risks of misuse and vulnerability for openly downloadable foundation models. The Partnership on AI (PAI) also recently released an excellent report outlining a list of risk mitigations organised by actors for (i) preventing, (ii) detecting, and (iii) responding to risks.[48] We highly recommend attending to the PAI report for a deeper dive into these options spaces. Here we list and augment some of the options from the PAI report that were also mentioned during the Open-Horizons workshop and that we find most promising (Table 3). We organise according to PAI's structure for consistency.

TABLE 3
KEY PLAYERS AND RISK MITIGATION THROUGHOUT THE AI VALUE CHAIN

PLAYERS	RISK MITIGATION
(i) Preventing Risk	
Model Providers	• Develop and implement durable model-level interventions (see 'technical solutions' above for examples).
Model Providers & Model Adapters	• Responsibly source and filter training data to reduce bias and remove harmful content. • Conduct internal safety and misuse evaluations to inform model release decisions. • Provide clear user guidance documentation.

[45] OpenMined (2024). Response to the National Telecommunications and Information Administration (NTIA) Request For Comment (RFC) on Dual Use Foundation Artificial Intelligence (AI) Models with Widely Available Model Weights. Retrieved July 12, 2024, from https://www.regulations.gov/comment/NTIA-2023-0009-0334
[46] See, for example, NVIDIA's Confidential Computing offering.
[47] Future of Life Institute (2023). Exploration of secure hardware solutions for safe AI deployment. Retrieved August 20 2024 from https://futureoflife.org/ai-policy/hardware-backed-compute-governance/
[48] Srikuman, M., Chang, J. & Chmielinski, K. (2024). Risk Mitigation Strategies for the Open Foundation Model Value Chain. Retrieved July 20, 2024, from https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

Model Hosting Services	• Establish consistent structures for content moderation on their platforms. • Assess whether hosted models meet the platform's standards for responsible model development and deployment including, for example, evidence of adequate safety testing and risk analysis, clear and complete documentation and model use guidance. • More closely monitor and focus evaluations on the most frequently downloaded models. While the open-source ecosystem is vast, 70% of hosted models have 0 downloads while 1% account for 99% of downloads thus narrowing down "widely used models" to a more manageable range.[49]
(ii) Detecting Risk	
Model Providers & Model Hosting Services	• Implement and support incident reporting channels to allow external stakeholders to report safety concerns, vulnerabilities and AI incidents. • Establish external audit and evaluation programs to facilitate access for auditors to critical components for detecting risk.
(iii) Responding to Risk	
Model Providers, Model Hosting Services, & App Providers	• Establish decommissioning and incident response policies outlining the conditions under which a model is recalled and no longer hosted, or changes to licence are implemented to limit or prohibit certain uses.

2.2.3 Mitigating risk through more openness

AI openness – mainly making AI Artefacts (e.g. code and weights) publicly downloadable – pose risks that can sometimes be reduced by more openness, or more specifically, by being better at being open and being more open about more things.

For example, better documentation (e.g. technical reports, model cards, and data cards with information model characteristics, training, and evaluation processes) and standardised documentation practices across platforms can enhance transparency and facilitate easier understanding and responsible use of AI artefacts. Open audits allow for independent verification of model safety and performance claims, fostering trust and accountability in the AI ecosystem. The availability of open datasets enables researchers and developers to train and test models on well-understood, ethically sourced data. Open benchmarks provide standardised ways to evaluate and compare different models, promoting fair competition and progress tracking. Being open about guardrails – the safety measures and ethical constraints implemented throughout the AI value chain – allows for collaborative improvement of these critical protective features.

2.2.4 Mitigating Risk through Systemic AI Safety

Finally risks from open-source model sharing can also be reduced by improving systemic AI safety. The systemic AI safety approach involves developing comprehensive strategies to create an environment where AI's potential for societal harm is inherently lower due to improved social

[49] Osborne, C., Ding, J., & Kirk, H. R. (2024). The AI Community Building the Future? A Quantitative Analysis of Development Activity on Hugging Face Hub. Retrieved July 22, 2024, from https://arxiv.org/abs/2405.13058

resilience and fortified systems.[50] For example, disinformation risks and the associated threats AI pose to democracy can be mitigated in part through enhanced digital literacy. They can also be mitigated by working to address the fundamental roots of distrust in politicians and democratic institutions that make the public more vulnerable to information manipulation.[51] By fostering a well-informed populace with strong critical thinking skills, we can create a more resilient society better equipped to handle the challenges posed by widely accessible AI technologies. This can also be achieved by providing tools to help with information evaluation (e.g. content provenance technologies), and by building a democratic system more deserving of trust as politicians and policy are seen to respond to citizen needs.

To be very clear, we are not suggesting that the onus should be on society to just deal with AI threats better. Rather, we are recognizing that AI impacts do not exist in isolation, but are a result of AI capabilities and misuse meeting the realities of the social systems in which the technology is embedded. Restricting access to models with dangerous capabilities and minimising opportunities for misuse are only part of what we can do to keep ourselves safe.

2.3 INSIGHTS ON PURSUING OPENNESS BENEFITS WHEN ACCESS RESTRICTIONS ARE IN PLACE

Even with working to reduce the risks of openness, some cases may still arise in which model access will be restricted. This may be done because the level of risk posed is still too high for responsible open release, or because model developers wish to keep their models private out of proprietary concern. In either case, we might lessen the hit to the benefits of openness by pursuing the benefits in other ways.

The following options are not perfect substitutions for open-source model release. Rather, the idea is that if we can be specific about which benefits we wish to pursue, we might be able to identify other strategies that go some distance in pursuit of those goals.

Here we summarise options highlighted during the workshop.

Some of options, marked (*), are alternative model access options that try to convey some benefits of openness while not making the model publicly downloadable.

Other options, marked (**), are things that can be done alongside model release - irrespective of how the model is released - to promote openness benefits.

We have roughly organised the options by the openness benefits they promote. An extended discussion on many of these points is presented in Seger et al. (2023).[52]

2.3.1 Facilitating external oversight

*****Privileged audit access:** Grant privileged model access to trusted, independently selected, third-party auditors via gated-download or, more securely, research API. While research APIs are not yet fully realised, there is hope that suitable access for auditors could be provided via a structured access approach.[53] For example, as part of the Christchurch Call Initiative on Algorithmic Outcomes, OpenMined collaborated with LinkedIn and Dailymotion to pilot Syft, which makes it possible to remotely study sensitive datasets and proprietary algorithms without accessing or compromising the

[50] AISI UK (2024). Systemic AI Safety Fast Grants. Retrieved July 12, 2024, from https://www.aisi.gov.uk/grants

[51] Seger, E. (2023). Generative AI and Democracy: Impacts and Interventions. Retrieved July 12, 2024. from https://demos.co.uk/research/generative-ai-and-democracy-impacts-and-interventions/

[52] Seger, E. et al. (2023). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Retrieved July 20, 2024, from https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[53] Bucknall, B. & Trager, R. (2023). Structured access for third-party research on frontier AI models: Investigating researcher' model access requirements. Retrieved July 12, 2024, from https://oms-www.files.svdcdn.com/production/downloads/academic/Investigating_Researchers%E2%80%99_Model_Access_Oct23-compressed_3.pdf

security of the underlying data using open-source tools.[54] Model owners load information pertaining to a model into a server, and then external researchers and auditors can query the servers and extract answers without gaining access to the underpinning data.

***Red-team community:** Establish a community of independently selected and pre-vetted red-team professionals to stress-test models pre-release. Members of the red-team community are provided gated access to the model for study instead of making the model fully publicly available. The community should be composed of members from the wider AI community as well as security professionals, and representatives from high-risk domains to which foundation models might be put to use. This is a promising role for a network of AI safety institutes to play.

***Staged release:** Incrementally release larger and more capable model versions. After each release, take time to evaluate how the models are being used and the societal impacts to inform if and how the next model version should be released.[55]

****Safety bounties:** Developers or governments establish safety bounty programs analogous to bug bounty programs commonly used in cybersecurity.[56] Bug bounty programs would offer financial and reputational rewards to members of the public who discover and responsibly report new safety failures, such as novel jailbreaks, or capabilities beyond those found in internal tests.

2.3.2 Facilitating beneficial AI progress (safety & capability research / new application development)

***Researcher access programmes:** Platforms can invest in developing researcher access programs that provide researchers secure access to data and models via research API and privacy preserving access control mechanisms.[57,58] In addition to information access, researcher access programs should also consider broader researcher needs including community support, network building, and resource investments. OpenMined is currently developing one such researcher access program with Reddit.[59]

****Profit commitment:** Companies commit a certain percentage of profits or research hours towards AI safety projects or social benefit research to drive progress in these directions.

****Incentive structures:** Governments can build incentive structures like large rewards programs for major scientific discoveries (e.g., protein folding) or pro-social advances (e.g. health and equity applications) using AI and for AI safety breakthroughs (e.g., interpretability).

***Plugins:** New integrations and applications can be explored and implemented through the development of "plugins" that allow a model to integrate with other services. The plugin could be submitted to the developer or a third-party auditor before publication. This option provides a mechanism for new integrations and applications to be reviewed and approved before being shipped while still tapping into public creativity and representation of interests and needs. A shortcoming is that downstream developers have less insight to study and test the safety of their integrations themselves.

***KYC gated access:** Developers can offer gated model access (i.e. full download access restricted to identified third parties) coupled with Know-Your-Customer (KYC) Requirements.[60] KYC requires

[54] OpenMined: Privacy-preserving third-party audits on Unreleased Digital Assets with PySyft. Retrieved August 18, 2024, from https://www.gov.uk/ai-assurance-techniques/openmined-privacy-preserving-third-party-audits-on-unreleased-digital-assets-with-pysyft

[55] Solaiman, I. et al. (2019). Release Strategies and the Social Impacts of Language Models. Retrieved July 15, 2024, from https://arxiv.org/abs/1908.09203

[56] Levermore, P. (2023). AI Safety Bounties. Retrieved July 15, 2024, from https://rethinkpriorities.org/publications/ai-safety-bounties

[57] Bucknall, B. & Trager, R. (2023). Structured access for third-party research on frontier AI models: Investigating researcher' model access requirements. Retrieved July 12, 2024, from https://oms-www.files.svdcdn.com/production/downloads/academic/Investigating_Researchers%E2%80%99_Model_Access_Oct23-compressed_3.pdf

[58] OpenMined: Privacy-preserving third-party audits on Unreleased Digital Assets with PySyft. Retrieved August 18, 2024, from https://www.gov.uk/ai-assurance-techniques/openmined-privacy-preserving-third-party-audits-on-unreleased-digital-assets-with-pysyft

[59] Reddit (2024). Publishing Our Public Content Policy and Introducing a New Subreddit for Researchers. Retrieved August 21, 2024, from https://www.redditinc.com/blog/publishing-our-public-content-policy-and-introducing-a-new-community-for-researchers

[60] Anderljung, M. et al. (2023). Frontier AI Regulation: Managing Emerging Risks to Public Safety. Retrieved July 15, 2024, from https://arxiv.org/abs/2307.03718

developers to vet and keep a record of model recipients. More so than plugins, the ability to download the full model allows downstream developers to more thoroughly understand and test the safety and performance of their integration while a mechanism for oversight is still maintained. A shortcoming to take into consideration is that models shared in this way are more likely to be leaked.

2.3.3 Improving competition, reducing market and power concentration, and mitigating vendor lock-in

*/****Interoperability requirements:** Governments or regulators could enforce interoperability requirements that involve standardising data formats and APIs. This will make it easier to transfer data and model outputs between different AI systems, reducing dependence on a single vendor's proprietary data formats. Standardised interfaces and protocols also allow for easier integration of AI models from different vendors into existing systems allowing downstream developers to mix and match solutions from various providers.

****National AI models:** Government could invest in building a national, publicly owned foundation model.[61] Several governments globally are currently working on sovereign models, including India, Singapore, and Taiwan. The aim is not necessarily to compete with frontier AI developers, but is primarily to focus on solving market failures by building "AI for Good" - safe AI systems made freely accessible to UK businesses to build upon and integrate into products, and to underpin AI applications in UK public services. The aim is to reduce public sector and SME reliance on big tech AI developers thereby shifting power, improving competition, and enabling digital autonomy from big private providers.

****Public compute infrastructure:** Compute (the processing power required to train and run AI models) is vital for the development and deployment of advanced AI models. Currently most compute resource is provided by private 'hyperscalers' such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud. By investing further in publicly accessible compute infrastructure and making that resource available for free or reduced cost to startup and SME, governments can help to overcome the capital investment required for new companies to break into a highly consolidated market. This could also help to reduce reliance on private providers, and provide secure, domestic infrastructure for processing sensitive national data.[62] Additionally, the public sector is arguably well-placed to deliver this infrastructural development, given that it has higher risk-bearing and coordinating capacity, and lower borrowing costs than private investors.[63]

****Open data ecosystems:** Governments and organisations working in data can contribute to and enforce open-data sets and ecosystems that allow as many people and communities as possible to access the data sets and data infrastructure they need to conduct research and to train or fine-tune new AI models.[64] Open data is particularly important for startups and SME looking to compete with big tech firms with links to vast data resources or means of data production. The maintenance of open or 'public-good' data sets - data sets made freely available for use by the general public, researchers, organisations, and businesses - can help balance this data asymmetry. Such work would push forward and build on ongoing efforts by, for example, data.go.uk[65] and Data Commons[66] and LAION.[67] Where fully open data sets raise privacy or security concerns (e.g. with health data), data sets might be made available through privacy preserving structured access approaches.[68,69]

[61] Belfield, H. (2023). Great British Cloud and BritGPT: The UK's AI Industrial Strategy Must Play to Our Strengths. Retrieved July 15, 2024, from https://www.labourlongterm.org/briefings/great-british-cloud-and-britgpt-the-uks-ai-industrial-strategy-must-play-to-our-strengths

[62] Lawrence, D. & Seger, E. (2024). GB Cloud: Building the UK's Compute Capacity. Retrieved 19 July, 2024, from https://demos.co.uk/research/gb-cloud-building-the-uks-compute-capacity/

[63] Shearer, E., Davies, M. and Lawrence, M. The Role of Public Compute. Ada Lovelace Institute. Retrieved 21 July 2024, from https://www.adalovelaceinstitute.org/blog/the-role-of-public-compute/

[64] ODI Policy Manifesto (2024). Retrieved July 10, 2024, from https://theodi.cdn.ngo/media/documents/ODI_Policy_Manifesto.pdf

[65] https://www.data.gov.uk/

[66] https://datacommons.org/

[67] https://laion.ai/

[68] OpenMined: Privacy-preserving third-party audits on Unreleased Digital Assets with PySyft. Retrieved August 18, 2024, from https://www.gov.uk/ai-assurance-techniques/openmined-privacy-preserving-third-party-audits-on-unreleased-digital-assets-with-pysyft

[69] OpenMined (2020). Privacy-Preserving Data Science Explained. Retrieved August 18, 2024, from https://blog.openmined.org/private-machine-learning-explained/

**__Democratic decision-making:__ Specifically to mitigate power concentration and autocratic control over AI, governments and AI developers might implement more democratic processes to guide complex and high-impact decisions about AI development, use, and governance, including decisions about model access. For example, participatory platforms such as Pol.is or Remesh might be used to synthesise public input to inform complex normative decisions about AI. Alternatively, representative deliberations, such as citizens assemblies can convene representatives of impacted populations to tackle AI governance questions.

Aside from directly eliciting public input, large AI developers could also adopt institutional structures to maintain transparency of internal processes and to dissipate control over high-impact decisions, even where proprietary model ownership is maintained. Options include implementing democratically selected oversight boards to vote on key issues, and incorporating as a public benefit company (PBC) to provide legal standing for prioritising public benefit over shareholder interests where public and shareholder interests collide.

SECTION 3
OPEN QUESTIONS ABOUT OPENNESS

We identified some open questions about AI openness that need more research to make meaningful policy progress.

3.1 WHAT ARE THE LIMITS OF THE ANALOGY BETWEEN OPEN-SOURCE SOFTWARE (OSS) AND OPEN-SOURCE AI?

The benefits of open-source have been largely built on the back of the software industry. Accordingly, the clearly evidenced successes of OSS are very often used to directly substantiate claims about the importance of maintaining open-source environments for AI development. There certainly is merit in the comparison, and there will be much overlap in the benefits. However it is not entirely clear that all of the benefits of open-source software development translate seamlessly to the context of AI, especially with respect to frontier model development. Both to lend credence to arguments in favour of open-source AI, and to protect ourselves in cases of disanalogy, the comparison needs clarity.

For example, in the case of OSS, it is generally well accepted that the offence-defence balance – a term referring to the "relative ease of carrying out and defending against attacks"[70] – most often comes out in favour of defence. Software vulnerabilities are relatively easy to find for developers and attackers, and software patches are relatively easy to make and usually fully resolve the vulnerability. There does, however, remain a challenge with patch adoption for OSS. A 2000-2018 survey of 150,000 medium and large U.S. organisations showed 57% of organisations using server software with known vulnerabilities even where more secure updated versions were available.[71]

In the context of AI the offence-defence balance is less well understood and there is a chance open-source publication of models may skew more towards offence than it does for OSS, especially for

[70] Garfinkel, B. & Dafoe, A. How does the offense-defense balance scale? Journal of Strategic Studies, 42(6):736–763, September 19, 2019. DOI: 10.1080/01402390.2019.1631810

[71] Murciano-Goroff, R., Zhuo, R. & Greenstein, S. (2024). Navigating Software Vulnerabilities: Eighteen Years of Evidence from Medium and Large U.S. Organizations. National Bureau of Economic Research, Cambridge, MA. Retrieved August 14, 2024 from doi: 10.3386/w32696.

more highly-capable models:[72,73] (i) Given our current lack of understanding of how advanced AI systems work internally, it may be difficult to identify the source of risk or failure; (ii) certain risks, such as bias and discrimination, may be learned from the training data, and it could be impossible to "remove" all bias from training data; (iii) reducing misuse of AI systems may require changes to social systems beyond changes to technical ones; (iv) the structure of AI systems introduces new sources of failure specific to AI that are resistant to quick fixes (e.g., the stochastic nature of large language models may make it difficult to eliminate all negative outputs, and the inability to distinguish prompt injections from "regular" inputs may make it difficult to defend against such attacks). These are preliminary thoughts, but worth investigating further.

Beyond the offence-defence balance, there are other axes along which the software-AI analogy can be evaluated. First there is a definitional question - how does the term "open-source" apply to AI- though clear and ongoing progress is being made on this front by the Open Source Initiative (OSI).[74] The analogy can also be evaluated with respect to accessibility (e.g. costs and expertise providing potentially higher barriers to entry) and how open-source software v. open-source AI will impact market concentration and how important it is for reducing profit concentration. The analogy will also likely carry over in different ways and to varying extents for different kinds of models of varying size and complexity.

3.2 WHAT IS THE IMPACT OF RESTRICTING ACCESS TO HIGHLY-CAPABLE AI MODELS ON COMPETITION AND MARKET CONCENTRATION?

As discussed in section 1.2, our group contained strongly differing opinions on the market implication of model access which we agreed stemmed at least in part from a lack of clear information as well as uncertaining about the trajectory of AI innovation.

There is an open question as to what the most financially lucrative and commercially viable AI systems will be in five, ten, and twenty years from now. Which models will be most useful, generate the most value, and be most economical to produce?

On the one hand, if despite rising training compute costs, larger training runs continue to yield massive capability gains, then the largest frontier foundation models may be the most economical as they can be extremely useful across a wide variety of applications. In such a case, restricting access to those frontier models could be detrimental to competition by preventing information sharing and allowing incumbents to more easily capture the market.

Another path toward a similar outcome when frontier models are generally capable and expensive to train is via 'economies of scope'. In economies of scope incumbents that provide a wide range of AI services are able to spread the costs of training foundations models and recoup their spending more easily compared to smaller firms that only develop a few applications.

On the other hand, if high training compute costs for the largest models combine with low marginal gains in capability, this may render smaller models and narrow AI applications more commercially viable. In such a case, restricting access to the largest, most highly-capable models may have a less detrimental impact on AI marketplace competition as most AI benefits and commercial activity would be taking place behind the frontier.[75]

There are also open questions, as discussed in section 1.2, about how consumers decide which AI models to use. For example, why do some governments choose to procure closed-source solutions

[72] Seger, E. et al. (2023). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. Retrieved July 20, 2024, from https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[73] Shevlane, T. & Dafoe, A. (2020). The Offense-Defense Balance of Scientific Knowledge: Does Publishing AI Research Reduce Misuse? Retrieved August 14, 2024 from https://arxiv.org/abs/2001.00463

[74] OSI (2024). The Open Source AI Definition - Draft 0.0.8. Retrieved 20 July, 2024, from https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8

[75] The United Kingdom's Competition and Markets Authority's (CMA) 2023 report outlines in detail AI market dynamics and sources of uncertainties such as this.

to support the public sector instead of building on open-source options? Under what conditions does building solutions on top of open models instead of procuring closed solutions (or vice versa) become economically preferable?

In order for governments to be well-informed about the economic implications of policy regarding AI openness and model restrictions, much more focused research is needed. This research should include systematic quantitative and qualitative studies examining what actors opt to build on open options, quantifying the switching costs of moving between closed models vs. open models to better understand the dynamics of vendor lockin-in, and studying where open-source model are used, and the possible long term economic consequences of relying heavily on private closed systems across multiple industries.

3.3 AT WHAT POINTS CAN MORE DEMOCRATIC AI GOVERNANCE PROCESSES PRACTICALLY AND MOST USEFULLY BE IMPLEMENTED?

A key benefit of greater openness across the AI value chain is to disperse power contraction from around big tech and to reduce instances of unilateral decision-making by large AI developers that will have profound impacts on all society. As discussed in section 2.3.3, besides democratising access to the technology itself, an additional method for pursuing the desired power distribution is via the democratisation of AI governance decisions.[76] That is, to distribute influence over decisions about how AI is developed, used, deployed and distributed by introducing more democratic processes such as citizens assemblies or multi stakeholder consultation to guide those decision-making processes. The benefit would not only be reduced power concentration, but also to build societal trust that AI solutions are being built for and with people.

There is an open question, however, about where more democratic governance ought to come into play - where is it practicable, and where would it be most useful for distributing control away from big tech ensuring societal needs and values are met. No one would suggest, for instance, that the day-to-day decisions of individual AI engineers at Meta or OpenAI be dictated by democratic insight. But there is certainly room at higher levels of abstraction for determining what values AI should be aligned with (e.g. via alignment assemblies)[77] and for informing governance decisions about how public funds are invested around AI, what application spaces are prioritised, or defining acceptable and unacceptable use cases.

Another possibility is that foundation model development is facilitated by independent foundations.[78] The vendor neutrality and open governance facilitated by independent foundations have been proven to act as key structural enablers for collaboration, including between market rivals, on the development of foundational or "base-layer" open source software.[79] For example, Meta donated PyTorch - a highly successful machine learning software project - to the Linux Foundation.[80] In turn, the Linux Foundation established a governing board and technical steering committee to involve a diversity of stakeholders in project governance rather than Meta alone. However, it remains to be seen how this open governance model will translate to AI foundation model development, which involves the release of and collaboration on many components beyond software.[81] A number of questions are raised. What are the merits and limitations of the foundation-hosting model for the development of models, and how might this model need to be innovated to accommodate for the nuances of model development? What conditions (e.g., economic, social, or legal) would incentivize (competing) companies to donate models (as well as other components, such as data) to independent foundations?

[76] On the four meanings of AI democratisation - democratisation of use, development, profits, and governance - see Seger, E. et al. (2024). Democratising AI: Multiple Goals, Meanings, and Methods. https://dl.acm.org/doi/10.1145/3600211.3604693

[77] Collective Intelligence Project (2023). Alignment Assemblies. Retrieved July 19, 2024, from https://cip.org/alignmentassemblies

[78] Linux Foundation (2021). Understanding Open Governance Networks. Retrieved August 18, 2024, from https://www.linuxfoundation.org/blog/blog/understanding-open-governance-networks

[79] Germonprez, M. et al. (2013). Open-source communities of competitors. Retrieved August 18, 2024, from https://doi.org/10.1145/2527191

[80] Zemlin, J. (2022). Welcoming PyTorch to the Linux Foundation. The Linux Foundation. Retrieved August 16, 2024, from https://www.linuxfoundation.org/blog/blog/welcoming-pytorch-to-the-linux-foundation

[81] White et al. (2024). The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. Retrieved August 16, 2024, from https://arxiv.org/abs/2403.13784

More work is needed defining where democratic insights should inform AI governance and determining what methods for gathering and integrating those insights are appropriate for which contexts.

3.4 HOW DO THE RISKS AND BENEFITS OF AI OPENNESS MAP OUT ACROSS A FINE-GRAINED TAXONOMY OF DEGREES OF AI OPENNESS?

In general, the benefits of AI openness increase with more openness across the spectrum while the risks of openness are primarily posed by model release - see section 1.2. However, we need a much finer grained understanding of how different aspects of AI openness and combinations thereof translate to risks and benefits to guide well informed and effective policy.

For instance, releasing different combinations of model components will pose different risks and opportunities for downstream developers and users, and many AI experts hold this understanding as tacit knowledge. It would be an extremely helpful and low-hanging-fruit exercise for researchers to lay out the activities (and associated risks and opportunities) enabled by publicly releasing different combinations of model components in easily referenced detail. Model-sharing standards should both support safe model distribution and protect open-source practices and benefits. To achieve both, these standards must be fine-grained and built on a well-researched and precise understanding of the extent to which access to different combinations of model components enable unrestricted model use, reproduction, and modification.

3.5 BEYOND RISK OF MODEL MISUSE AND VULNERABILITY PROLIFERATION, WHAT ARE 3HE NATIONAL SECURITY IMPLICATIONS OF AI OPENNESS?

Discussion about the national security implications of AI openness usually revolves around the risks of model release - how downloadable models open a door for malicious actors to bypass safeguards and identify vulnerabilities that can be exploited wherever the model has been integrated in downstream applications. This is of significant concern where adversaries are looking to attack or undermine other state actors through malicious use or by exploiting vulnerabilities.

However the national security implications of AI openness are much broader than adversarial attack. Below we outline additional considerations pertaining both to positive and negative national security implications. It is not an exhaustive list and more research is needed to fully flesh out these implications.

National security risks of AI openness:

- Malicious use of open models by adversarial states or malicious actors to undermine national security.
- Exposure of vulnerabilities in models used in national security applications.
- Possible arms race acceleration and open sharing narrow the gap between AI leaders and laggers.
- Possible loss of strategic advantage (for AI leading nations with a strategic advantage to defend).

National security benefits of AI openness:

- Enables digital autonomy - Open AI tools reduce reliance on proprietary systems controlled by a few large tech companies or powerful nations. Open AI tools can also be used to process and analyse sensitive data locally, rather than relying on external services.

- Talent pool expansion - Open AI ecosystems may attract and nurture more diverse talent, strengthening a nation's overall AI capabilities.
- Facilitate International cooperation - greater openness could foster collaboration between allied nations, pooling resources and knowledge to address common security challenges.

3.6 WHAT GUARDRAILS CAN BE IMPLEMENTED ACROSS THE AI STACK TO REDUCE THE RISK OF AI OPENNESS?

As summarised in section 2.2.2, PAI has made a good start answering this question.[82] We recommend expanding upon the work to consider risk mitigation strategies outside the AI stack, looking also to ways in which risks of AI openness can be mitigated through improvements to systemic AI safety (section 2.2.4).

Systemic AI safety is about developing comprehensive strategies to create an environment where AI's potential for societal harm is inherently lower due to improved social resilience and fortified infrastructure.[83] Some examples of systemic safety project that could help mitigate risks from openness include:

- Investigating systems-based approaches to improving trust in authentic media and expert voices while mitigating the spread of AI-generated misinformation.
- Restoring trust in democracy and policy makers through more collaborative democracy[84] in order to mitigate negative impacts of AI-enabled influence operations.
- Investigating technical and policy methods for tracking and criminalising online abuse using deepfakes.
- Targeted interventions to protect critical infrastructure such as energy, finance, or healthcare infrastructure from AI-enabled cyberattack.

3.7 HOW CAN OR SHOULD LIABILITY BE SHARED AND TRANSFERRED BETWEEN MODEL PROVIDERS AND DOWNSTREAM DEVELOPERS/USERS?

Unclear liability can stifle open-source development. Model developers will be less inclined to release model versions for others to study and iterate on if they are worried about being held fully responsible for any harms coming from downstream applications of their works in progress. Downstream developers will also be less inclined to experiment with and build on open-source models if they too would be exposing themselves; liability is often assigned through contracts, and several workshop participants felt that this was placing too much risk on downstream developers and users of open foundation models who are often unable to effectively mitigate risks because of limited model transparency (e.g. insufficient access to documentation, testing records, training data etc.).

Our group largely agreed that regulation that clarifies how liability transfers with the release and use of open-source models would help support downstream developers and preserve open-source ways of working. One option is to make it a condition of transferring liability downstream to meet sufficiently high standards of openness, including sharing complete documentation such as technical reports, model cards, and data cards. A degree of shared liability between foundation model developers and downstream developer/users might be appropriate, especially for more highly capable models.

[82] Srikuman, M., Chang, J. & Chmielinski, K. (2024). Risk Mitigation Strategies for the Open Foundation Model Value Chain. Retrieved July 20, 2024, from https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/
[83] AISI UK (2024). Systemic AI Safety Fast Grants. Retrieved July 12, 2024, from https://www.aisi.gov.uk/grants
[84] Levin, M., Curtis, P., Castell, S. & Kapetanovic, H. (2024). Citizens' White Paper. Demos. Retrieved August 10, 2024, from https://demos.co.uk/research/citizens-white-paper/

SECTION 4
MENU OF OPTIONS FOR GOVERNMENT FOR PROMOTING THE BENEFITS OF OPENNESS WHILE MITIGATING RISKS

Building on the insights presented so far in this report, we have compiled the following list of policy options that the government might pursue to help maximise the benefits of AI openness for its citizens and AI industry while mitigating undue risk. For each policy option we provide a description of how it prompts openness benefits and a brief commentary on the potential shortcoming or challenges of the policy option to serve as a starting point for further investigation.

TABLE 4
POLICY OPTIONS FOR PROMOTING BENEFITS OF AI OPENNESS

POLICY OPTION	IMPLICATIONS	
4.1 INVESTMENTS		
1. Provide financial support for open-source projects and ecosystems.[85] *In addition to funding open-souce projects and developers, this might also include support for open-source safety testing ecosystems through the provision of compute resource, tools, and standards (e.g. AISI's open sourcing of Inspect)*[86]	Strengths	• There is potential to stimulate innovation, e.g., as more people might be willing to help if there is a clear public benefit goal. • Government funding can steer projects attention e.g. towards security, safety best practices, data curation, and other beneficial interventions. • Offers some protection against power concentration.

[85] See Milton, T., Osborne, C., & Pickering, M. (2024). A UK Open-Source Fund to Support Software Innovation and Maintenance. Retrieved August 16, 2024, from https://ukdayone.org/briefings/a-uk-open-source-fund. For examples of public-private open-source funding modes see Osborne, C. (2024). Public-private funding models in open source software development: A case study on scikit-learn. Retrieved August 10, 2024, from https://arxiv.org/html/2404.06484v1

[86] Gov.uk (2024). AI Safety Institute releases new AI safety evaluations platform. Retrieved August 15, 2024, from https://www.gov.uk/government/news/ai-safety-institute-releases-new-ai-safety-evaluations-platform

	Strengths	• Helps maintain the functionality and security of key open-source components • There is an opportunity to build models tailored to different cultures or sectors e.g., trained on non-English languages (Welsh, Gaelic, languages of migrant communities), tailored to disadvantaged communities, vulnerable people, etc. • Well-supported open-ecosystems could attract more talent to the country. • If thriving open-source ecosystems stimulate competition, then this could also attract more AI investment to the country.
	Weaknesses	• It is difficult to assess return on investment for supporting open-source projects. The success of individual open-source projects is particularly difficult to assess in advance.
2. Invest in the development of a national foundation model *Possibly through public private partnership* See: Section 2.3.3	Strengths	• Secure sovereign control over domestic AI capabilities. • Subsidised access can be provided for startups and SMEs to facilitate innovation and greater competition. • Supports a culture of collaboration to facilitate continued strengthening of the foundation model.
	Weaknesses	• An unchecked vulnerability in the foundation model could implicate all other models built on top (however, this is true of all models, not just sovereign models). • It is an expensive option. For example, the Spanish government is working in collaboration with IBM, who have received a very large (expensive) contract. Further, BLOOM had access to a French supercomputer, an important example of access to compute infrastructure. • We may not want all models to be open for national security reasons (e.g. military applications).

3. Invest in building and maintaining 'public good' data sets or 'Open Data Libraries' as stipulated in the Labour Manifesto.[87] See: Section 2.3.3	Strengths	• Open datasets create more opportunity for startups to innovate by lowering costs. • Open data sets could be curated by the community to reduce harmful and biassed content. • Open datasets could democratise access to AI by providing more parties with access to the quality data needed to train and fine-tune models. • Researchers and civil society are able to scrutinise open data sets to further ensure quality and to mitigate instances of models being trained on biassed or corrupt data.
	Weaknesses	• Opening sensitive data raises security and data privacy concerns. Structured transparency[88] methods might offer a middle ground solution to enable research and auditing without openly sharing data.
4. Public compute investment *The government can not realistically invest enough to compete with leading hyperscalers (e.g. AWS, Azure, Google Cloud). Therefore, the goal should not be to compete, but to support a nation's researchers, public services, and start-ups so as to meaningfully diversify compute supply.* See: Section 2.3.3	Strengths	• Government-owned compute is easier to regulate. • Diversifies compute supply so as not to rely exclusively on foreign hyperscalers and big tech companies. • Provides a local and secure compute option for processing private and sensitive public sector data. • Free or reduced cost access can be provided to start-ups, SMEs, and researchers to promote innovation.
	Weaknesses	• Some benefits of public compute might also be achieved through careful procurement processes.
5. Invest in research towards potential technical solutions for mitigating risks from open models (e.g. through large grant programs). See: Section 2.2.1	Strengths	• Research could identify technical options with better tradeoffs between closed and open AI models.
	Weaknesses	• Funding may not be the bottleneck for such research, but rather it is access to models, data or more talent that is needed.

[87] Change: Labour Party Manifesto 2024. Retrieved July 20, 2024, from https://labour.org.uk/wp-content/uploads/2024/06/Labour-Party-manifesto-2024.pdf

[88] OpenMined (2021). Structured Transparency: Ensuring Input and Output Privacy. Retrieved August 19, 2024 from https://blog.openmined.org/structured-transparency-input-output-privacy/

6. Invest in clear threat modelling exercises involving domain experts to underpin targeted risk mitigation policy. *For example, this could be tasked to AISI, who then build upon these threat models with investment in building domain specific risk mitigation strategies.* See: Section 2.1.2 and a similar recommendation made the the U.S. NTIA for "Developing and maintaining a set of risk portfolios, indicators, and thresholds."[89]	Strengths	- It makes the case clear to regulators what threats AI safety measures are trying to defend against. - Doing this process collaboratively could have good implications.[90] For example, this solution could help build a common understanding of AI ecosystems, facilitate the sharing of information relating to security threats (e.g. this could be modelled on cybersecurity databases, such as MITRE's ATT&CK, which is a globally accessible knowledge base of adversary tactics).
	Weaknesses	- Fully public threat modelling exercises could pose a potential infohazard, for example, it could focus the attention of malicious actors on the more dangerous threats. It may be necessary to keep some threat modelling secure. - There is no consensus on how threat modelling should be done. For example, what is standard practice? Who is involved? It's very difficult to get consensus on a threat model being properly done, complete, or accurate.
7. Investigate economic impacts of open-sourcing AI models and of restricting model access. *This might include, for example, research examining the trade-offs between allowing downstream developers to innovate using models they would not afford themselves, versus providing large developers with a mechanism for further entrenching their positions as industry leaders.* See: Section 1.2 and 3.2	Strengths	- We need to understand better how access restrictions on different kinds of AI models impact market concentration and competition. With this information governments can make much clearer and targeted decisions about how workings to reduce risks from AI openness interact with competition and market concentration concerns.
	Weaknesses	- It's difficult to project how AI capabilities will develop and where the market will settle - e.g. with large general use foundation models or with smaller, narrowing systems. - Adoption and adaptation is slow and evidence gathered today might be misleading with respect to tomorrow's AI technologies.

[89] NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

[90] See, e.g., Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213 (p.14).

8. Investigate guardrails that can be deployed throughout the AI value chain to reduce risks of AI openness. *Building on PAI's recent option space overview.*[91] See: Section 2.2.2	Strengths	- Deploying guardrains throughout the AI value chain provides more robust protection against risks. Restricting model release is an imperfect solution for preventing AI harms; leaks should be expected and some harms might be effectively mitigated by other means without infringing on the benefits of AI openness. - The use of guardrails throughout the lifecycle of AI model development and deployment makes it harder for one actor to bypass all guardrails alone.
	Weaknesses	- The nature of risks throughout the AI value chain can be uncertain. - Uncertainty also makes it difficult to detect and react to those risks. To do so might require a disproportionately high effort, whereas a 'light touch' approach might not catch enough risks.
9. Invest in incentive structures such as large rewards programs to promote AI safety and social benefit breakthroughs.[92] See: Section 2.3.2	Strengths	- The use of incentive structures could help to make AI safety the norm, if all contributors/users of the AI model want to adhere to guidelines.
	Weaknesses	- It is difficult to know in advance what impact (including potentially negative impacts) e.g., major scientific discoveries using AI might have. For example, scientific breakthroughs that could provide social benefit might also be used maliciously.
4.2 REGULATION		
10. Establish transparency requirements for highly-capable proprietary models. *As in Article 13 of the EU AI Act*[93]	Strengths	- Allows downstream developers to responsibly integrate potentially high-risk systems into new applications. - Enables more thorough safety testing at different stages along the AI value chain.[94] - Open-source models will already satisfy these requirements.

[91] Srikuman, M., Chang, J. & Chmielinski, K. (2024). Risk Mitigation Strategies for the Open Foundation Model Value Chain. Retrieved July 20, 2024, from https://partnershiponai.org/resource/risk-mitigation-strategies-for-the-open-foundation-model-value-chain/

[92] NSF funding for mechanistic interpretability research on open-models provides one such example. https://www.khoury.northeastern.edu/research_projects/national-deep-inference-fabric-ndif/

[93] EU AI Act. Article 13: Transparency and Provision of Information to Deployers. https://artificialintelligenceact.eu/article/13/#:~:text=This%20article%20states%20that%20high,limitations%2C%20and%20any%20potential%20risks.

[94] Along similar lines, the U.S. NTIA report recommends that the US government might compel auditing and transparency for closed weight foundation models to enable independent government evaluations. NTIA Report (July 2024). Dual-Use Foundation Models with Widely Available Model Weights. Retrieved August 10, 2024, from https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

	Weaknesses	• Compliance costs may be less manageable for smaller developers.
11. Introduce exemptions from AI regulation for models that meet a certain standard of openness/ transparency. *Such as in the EU AI Act article 2(12).*[95] *The exemption does not apply to models classified high-risk.*[96]	Strengths	• Incentivizes greater transparency and information sharing among AI developers • Smaller open-source developers and dispersed open-source communities are not burdened by regulatory requirements that they do not have resources or coordination to action. • Higher-risk systems can be bracketed off to not qualify for exemption irrespective of openness (e.g. 'high-risk' AI in the EU AI Act), though that category of technology will need to be carefully defined so as to protect against harms while minimising negative impacts on innovation.
	Weaknesses	• If regulatory exemptions are used to encourage greater openness but the transparency requirements for earning the incentive are too stringent, some providers who might otherwise have offered semi-open access to their models may be motivated to close model access. The perfect could become the enemy of the good. • If the exemption is applied too liberally, some potentially harmful systems may be exempt from regulation.
12. Clarify liability legislation and establish openness standards as conditions for transferring (some degree of) liability downstream. See: Section 3.7	Strengths	• Incentivise more open-source development by providing developers with greater certainty about their responsibilities and the conditions under which they would be held liable. • Provides downstream users with clear avenues for redress for AI harms.

[95] EU AI Act. Article 2: Scope. Retrieved 18 August, 2024, from https://artificialintelligenceact.eu/article/2/

[96] EU AI Act. Article 6: Classification Rules for High-Risk AI Systems. Retrieved 18 August, 2024, from https://artificialintelligenceact.eu/article/6/

	Weaknesses	• Clear legislation that places too much responsibility on open-source developers for downstream harms may encourage more closedness around model development which may ultimately be worse for AI safety. Liability sharing between upstream developers and downstream developers/distributors/users must be carefully balanced to facilitate innovation while incentivizing meeting high safety standards.
13. Define part of UK AISI's role as providing safety evaluation support for UK startup and open-source ecosystems - e.g. by providing resources, tools, or safety evaluation standards and services.	Strengths	• Supports AI developer who may have few resources and internal expertise to ensure they are deploying new AI tools safely and responsibly. • Wider communities of AI developers can offer feedback and help improve the evaluation services AISI provides.
	Weaknesses	• Need to be very clear about how AISI's "stamp of approval" conveys any protections from liability. Safety failures in systems that have passed AISI evaluations could reduce trust in government serving as an AI evaluator. • Open and consistent safety evaluation standards allow developers to train "to the test" and pass evaluations in narrow domains while demonstrating reduced performance in real world application.
14. Consider mechanisms for incentivising or enforcing system interoperability requirements. See: Section 1.2 and 2.3.3 *Interoperability describes the ability of computer systems or software to exchange and make use of information that involve standardising data formats and APIs. Interoperability could be enforced for proprietary models through licensing conditions and underpinned legal requirement, such as through antitrust or consumer protection law.*	Strengths	• This could make it easier to transfer data and model outputs between different AI systems, reducing dependence on a single vendor's proprietary data formats. • Standardised interfaces and protocols allow for easier integration of AI models from different vendors into existing systems, allowing downstream developers to mix and match solutions from various providers. • Comparing model performance and characteristics is easier for external evaluators when comparing interoperable models.

	Weaknesses	- There could be a stifling effect on innovation if interoperability requirements hinder developers' ability to explore new architectures and approaches that don't fit with requirements. Accordingly, perhaps interoperability should only be encouraged in some domains. - Compliance costs for interoperability requirements could be high and more easily managed by more well-resourced actors. - Standardised interfaces and protocols could yield common vulnerabilities to attack across multiple systems. - There is a very high bar to enforce interoperability requirement on the basis of consumer protection - harms to the consumer needs to be evidenced - so instances in which government action is justified may be limited.
15. Set public sector procurement standards for model transparency to incentivise greater openness around private models. *This might include, for example, requiring transparency around safety evaluation findings, or certain standards of interoperability in order for a model to be considered by procurement to a public sector application.*	Strengths	- Strongly incentivize providers to meet openness standards in order to acquire the government as a customer. - There is an opportunity for openness standards to proliferate throughout the AI ecosystem (e.g. analogous to the 'Brussels effect' - the process of unilateral regulatory globalisation caused by the European Union). - Greater openness of procured technology may reduce issues of provider lock-in allowing the government to more easily move to new providers if preferable.
	Weaknesses	- Too much openness could display vulnerabilities in procured models. - There will be a need to specify which aspects of openness are desirable (e.g. data sets, safety audits, etc.) and in which contexts.

16. Establish / reform government open data policy. *To make data usable and accessible across the economy, consider mandating that data be open by default, with organisations publishing reasons for not opening data.[97] Where data privacy and security concerns exist, structured transparency[98] methods might be offered a middle ground solution to enable research and auditing without openly sharing data.* See: Section 2.3.3	Strengths	• Provides startups and SMEs access to valuable data resources, otherwise only held by data producers, to enable model training. • Enables better use of data by local jurisdiction to seek insights about local challenges and formulate more effective responses.
	Weaknesses	• Too much openness around sensitive data raises security and data privacy concerns. Access considerations must balance individual rights and public benefit.
17. Incorporate democratic processes into government decisions around AI. *This could include, for example, decisions around spending and public service integration, and around access to AI models.* *Practical options might include e.g., implementing democratically selected oversight boards, and employing participatory processes facilitated by civic tech (e.g. platforms such as Polis and Remesh) to engage diverse multistakeholder deliberation* See: Section 2.3.3 and 3.3	Strengths	• Provides a mechanism for opening up and democratising AI governance. independently of model release decisions. • This option could increase public trust in AI models, and in government decisions around AI. • The employment of more democratic processes provides the government with accountability for its decisions.
	Weaknesses	• Many decisions about AI - e.g. individual coding decisions - do not lend themselves to wide public engagement. Research is needed to establish a taxonomy of AI governance decisions that would benefit from wider deliberative engagement. • Introducing democratic processes could introduce unnecessary bureaucracy to some decisions.

[97] See principle 2 of the ODI Policy Manifesto (2024). Retrieved July 10, 2024, from https://theodi.cdn.ngo/media/documents/ODI_Policy_Manifesto.pdf

[98] OpenMined (2021). Structured Transparency: Ensuring Input and Output Privacy. Retrieved August 19, 2024 from https://blog.openmined.org/structured-transparency-input-output-privacy/

CONCLUSION

Greater openness around AI is a worthy goal. Transparency, information sharing, and open-source development enables collaboration, fuels innovation, drives healthy competition, and improves AI quality and safety through community oversight. However there are also risks that come with greater openness including the risk of exposing vulnerabilities and sensitive information and opening a door for easier model misuse.

Balancing these risks and benefits is a persistent challenge. In this report we have reflected on the state of the AI openness discourse, noting continued areas of disagreement and emerging points of consensus as we chart a path forward. We have focused on identifying open questions about AI openness that require further investigation to progress areas of stagnated debate and to start pursuing solutions even where some disagreement may still persist.

We encourage the government to further explore the policy options we outline above. These options provide opportunities for pursuing the wide benefits of AI openness, reducing openness risks, and engaging alternative strategies for pursuing openness benefits where risks cannot be satisfactorily mitigated.

Much work is still needed to articulate safety evaluation standards and the clear threat models to underpin consistent model sharing guidelines. However, there is reason for optimism in the trajectory these discussions are taking. Many of the key questions we need to answer have been identified. Further, what has often felt like an enduring debate pitting risks against benefits is now progressing towards creative technical and policy strategies for harnessing the benefits of AI openness in all its forms.

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

> i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;
>
> ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS SEPTEMBER 2024

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK