R: 2024:10

Generative AI’s open source challenge:

policy options to balance the risks and benefits of openness in AI regulation

Nick Botton & Mathias Vermeulen (AWO)
October 2024

Commissioned by the Digital Infrastructure Insights Fund (D//F).

DIGITAL
/NFRASTRUCTURE
/NSIGHTS
FUND

Contents

Executive Summary

A key trend in the development of Generative AI models is openness, whereby developers make their models, model elements and associated information freely available to the public. Openness in the development of Generative AI models brings numerous benefits, including the distribution of knowledge and enabling a diverse set of actors to use and repurpose the technology. It can also enhance competition by enabling more organisations, including start-ups and small businesses, to build upon and integrate advanced technology, originally developed by large developers, into their own products and services. Openness, however, also carries risks, notably associated with malicious actors' ability to misuse the models in question, and remove the safeguards put in place by developers. Additionally, openness remains a complex and often misunderstood concept, allowing some companies to misrepresent the true extent of their model's openness —a practice increasingly known as "open washing". Together, these three developments constitute the **openness challenge**.

In order to preserve the advantages of openness, suggestions have been made to relax safety obligations for "open-source" models, in order to encourage experimentation and research, and to support economic growth. However, this paper argues that creating broad legal exemptions in legislative frameworks on AI is neither feasible (in the absence of a universally adopted definition) nor advisable for strengthening these benefits, especially considering the existing risks associated with openness. Given the field's rapid growth and innovation, it's highly likely that new risks will surface, extending well beyond those outlined in this paper. These factors make an explicit carve-out currently too blunt of an instrument, with significant potential for abuse and a disproportionate negative societal impact.

Rather than broad legal exemptions, this paper suggests several policy options designed to balance the risks and benefits of openness. This approach aims to foster a level of openness that supports the democratic governance of Generative AI while mitigating associated risks. The paper highlights the value of expanding access for external researchers to enhance model safety, coupled with measures to limit access for potential malicious actors. It argues that any AI policy framework that seeks to balance the risks and benefits of openness of Generative AI should enhance external researchers' role in model development and maintenance.

Currently, neither EU, UK nor US frameworks adequately tackle the issue of openness in Generative AI models. The frameworks either contain gaps in scope, are missing important provisions, or are voluntary in nature. Yet, the

EU, UK and US initiatives together provide an outline for what a framework that adequately tackles the issue of openness could look like. Such a framework would include elements such as the EU AI Act's focus on highly capable models and systemic risk, the EU DSA's framework for enabling external research and vetting researchers, the UK's focus on pre-release evaluations, and the US's focus on standards for red teaming and involving third parties in model development.

In the future, policymakers around the world should ensure that AI legislation tackles the openness challenge. In particular, the EU should take advantage of the opportunity to tackle this issue presented by the AI Act's Code of Practice for General Purpose AI, which could address how and when external researchers should be involved in risk evaluation and mitigation.

Key findings

1. Lack of clarity regarding what constitutes "open source" in Generative AI has resulted in open washing

Due to the ambiguous nature of what "open-source AI" exactly entails, Generative AI companies often use the term "open-source" in ways that diverge from its traditional meaning. Companies like Meta or OpenAI have used the words "open" or "open-source", but effectively mean different things. Furthermore, some of these entities impose restrictions on the use, reproduction, or modification of their models, actions that conflict with the core principles of open-source software. This practice, known as "open washing", involves companies branding their models as "open-source" or "open" as a form of misleading virtue signalling. Using open washing in this way undermines the public's understanding of AI, create diversions from the risks associated with Gen AI, and fosters a culture of openness that falls short of true transparency.

2. Open washing disproportionately focuses on promoting the benefits of openness without fully addressing its risks

Open washing, a tactic often used by organisations trying to limit AI regulation as much as possible in general, strategically highlights the benefits of open-source Generative AI while neglecting its risks. This practice complicates efforts to effectively regulate Generative AI, as these companies are keen to take advantage of AI regulations that put a lighter burden on broadly defined open-source models. When misused, certain forms of openness can enable and exacerbate online safety risks, because openness (1) allows malicious actors to evade developer oversight; (2) allows those actors to remove any safeguards built into models; and (3) allows models to be customised for harmful purposes.

3. Opening up access to external parties can improve risk mitigation measures

Openness has a double-edged impact on safety. While openness increases the ability of malicious actors to misuse the model, it also enhances the ability of external researchers to scrutinise these models, enabling them to identify risks and improve mitigations. To make the best of this benefit in tackling the openness challenge, it is important for policymakers to understand that openness is neither binary nor straightforward, but instead a complex spectrum: the parts of Generative AI models that developers make available vary significantly, and the methods they use to make those parts available varies as well.

Blanket legislative exemptions for “open” models therefore risk having unintended consequences, as the degree of openness of a model affects the balance between the benefits and risks it brings.

4. An open science approach to releasing models can lead to increased safety

Openness plays a pivotal role in revealing the intricacies of Generative AI models, enabling scrutiny that extends beyond developers’ own evaluations. When models are made available for download, users can interact with them without the limitations imposed by query API access, which typically restricts the extent of user interaction. However, full access to all model elements is not always necessary for achieving the safety benefits associated with openness. There exists a sweet spot in the degrees of openness of Generative AI models that allows developers to balance the benefits and risks effectively. This optimal point enables sufficient openness for external researchers to enhance safety, while restricting access to potential misusers.

Moreover, fostering openness towards external researchers is essential for capturing the full benefits of transparency in AI development. This inclusive approach should be encouraged both before and after a model’s release, irrespective of the eventual level of public access. Engaging a diverse range of experts throughout the development and post-release phases allows the AI community to better manage the complex interplay of risks and advantages these models present. Openness to external researchers can thus promote the democratic governance of Generative AI, help make individual models become safer, boost innovation in the development of safeguards through an open science approach, and drive the development of safety norms in the long-term.

5. Barriers exist that limit the potential of openness to external researchers

There are nonetheless several limitations that hinder the benefits associated with openness towards external researchers. Developers can pick and choose the areas that their contracted external researchers focus on, which may result in certain risk areas being ignored. Developers may also limit the types of information and the model elements that they make available to external researchers, further limiting the scope and robustness of research. Finally, the safe harbours linked to independent research are either lacking or imperfect, meaning that independent researchers may have their accounts suspended for attempting to make models produce content that is incompatible with the developers' policy. These limitations significantly restrict the effectiveness of openness to external researchers, and should be a priority for policymakers. They represent major barriers in achieving an optimal balance between the risks and benefits associated with the openness of Generative AI.

6. Current policy approaches do not adequately tackle the openness challenge

Neither EU, UK nor US frameworks adequately tackle the openness challenge of Generative AI models. Their frameworks either do not apply to Generative AI, are missing important provisions, or are voluntary in nature. Yet, the EU, UK and US initiatives together provide an outline for what a framework that adequately tackles the challenge of openness could look like. Such a model could include elements of the EU AI Act's focus on highly capable models and systemic risk, and initiatives facilitating external researcher access for audit and evaluations, while drawing on the UK's focus on pre-release evaluations, the US's focus on standards for red teaming and involving third-parties in model development, and the EU DSA's framework for vetting researchers and enabling access to a range of public and non-public data.

Recommendations

Policymakers around the world will have many opportunities in the immediate future to consider the challenge of openness. In the EU, the AI Act comes with numerous forms of secondary legislation which may respond to this issue. The UK is expected to introduce AI legislation in the future, and the US may do so as well. There are furthermore many governments around the world which are yet to develop their own AI policy frameworks. These discussions provide an opportunity to develop a nuanced approach to tackling the openness challenge in a way that preserves its benefits while fostering a level of openness that makes Generative AI models safer, irrespective of whether they are eventually made open source. This paper ends with a set of non-mutually exclusive policy recommendations that could be taken on board by countries who are drafting and/or reviewing their own AI legislation.

1 Introduction

One of the key challenges associated with Generative AI is openness, meaning developers make their models, model elements and associated information available to the public. Three dynamics characterise openness in the context of Generative AI: (1) **openness comes with benefits** associated with distributing knowledge and the ability to use and repurpose technology; (2) **openness comes with risks** associated with malicious actors' ability to misuse models outside of developers' control; and (3) openness in the context of Generative AI is currently poorly defined in existing legislation, allowing companies to market their models as "open" while misrepresenting how open their models actually are – a phenomenon called "**open washing**." Taken together, these three dynamics constitute the **openness challenge**.

The aim of this paper is not to challenge efforts to define open-source AI[1], which stipulate the elements that should be made openly available for a model to be called "open-source". Instead, this paper aims to highlight how policymakers can foster a level of openness that makes Generative AI models safer, irrespective of whether they are eventually made open-source or not. It highlights the benefits and ways in which openness to external researchers can help reduce the risks of Generative AI models. Additionally, this paper aims to link the conversations around openness to those on external evaluations of Generative AI models and highlight policymakers' options for addressing the challenge of openness in this context.

1.1 Structure

This paper is structured as follows:

1. **Introduction:** Introduces the openness challenge in connection with the risks and benefits of Generative AI, and open washing.
2. **The risks of openness**: Describes the ways in which Generative AI can amplify online safety risks, and how openness contributes to those risks.
3. **The impact of degrees of openness on safety risks:** Describes the different degrees of openness, and how they affect the risks described in section 2.
4. **Finding a balance through openness to external researchers**: Describes openness to external researchers as tool for democratising AI

[1] The Open Source AI Definition – 1.0. *Open Source Initiative.* https://opensource.org/ai/open-source-ai-definition

governance that balances the risks and benefits of openness of Generative AI models.

5. **Current policy approaches to openness**: Describes the extent to which the policy frameworks of the EU, US and UK tackle the openness challenge.
6. **Policy options to balance the risks and benefits of openness in AI regulation:** Describes a set of suggestions for balancing the risks and benefits of the openness of Generative AI models through legislation.

1.2 Problem definition: open washing and the risks and benefits of openness

When discussing the openness of their models, many developers label their models as "open-source" in the same way the term is often used in the context of traditional software. The Open Source Initiative, the most authoritative voice on open source, states that software can be called "open-source" if it is freely distributed along with its source code, allowing the free modification and recreation of the software, the creation of derived works, and with the absence of use limitations and restrictive licenses[2].

Open source is both a philosophy and a praxis. It sees software engineers freely share the software and components they have developed, contribute to each other's works, and propose solutions to shared problems. As a result, open source is an inherent part of the digital economy: virtually all software contains open-source code[3], and 70-90% of all code is open source[4]. OpenForum Europe (OFE) estimates that each 10% increase in contributions to open-source software could result in an increase of 0.4-0.6% in European GDP[5]. OFE estimates that the economic impact of open source in the EU ranges between €65 and €95 billion.

Accordingly, there are therefore significant economic benefits associated with making Generative AI models more open. Making models more accessible and usable means that more businesses can use and implement Generative AI in their services at zero cost, in theory distributing the economic benefits of the technology. This could have a positive impact on market concentration, by allowing more companies to create AI-powered services that compete with those of large developers. Additional benefits, as described in section 4 of this paper, include improving the ability of external researchers to scrutinise model flaws, and making the governance of AI more democratic.

In the realm of Generative AI, the term "open-source" is often used in ways that diverge from its traditional meaning and the principles upheld by the Open Source Initiative and its new Open Source AI definition[6]. Developers like Meta

[2] Open Source Initiative (2024). The Open Source Definition. https://opensource.org/osd
[3] Bals, F. (2024). "2024 Open Source Security and Risk Analysis Report." *Synopsis.* https://www.synopsys.com/blogs/software-security/open-source-trends-ossra-report.html
[4] Perlow, J. (2022). "A Summary of Census II: Open Source Software Application Libraries the World Depends On." *Linux Foundation.* https://www.linuxfoundation.org/blog/blog/a-summary-of-census-ii-open-source-software-application-libraries-the-world-depends-on
[5] OpenForum Europe (2021). Study about the impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy. *European Commission.* https://digital-strategy.ec.europa.eu/en/library/study-about-impact-open-source-software-and-hardware-technological-independence-competitiveness-and
[6] The Open Source AI Definition – 1.0. *Open Source Initiative.* https://opensource.org/ai/open-source-ai-definition

have championed openness as a force for good, releasing highly capable models for public download, promoting accessibility and collaboration[7]. Conversely, OpenAI ceased releasing models for download in 2019 due to concerns about potential misuse, although they still provide model access through an API. Mistral AI has also made models available for download but has stopped disclosing specific details about their training processes, citing the "highly competitive nature of the field"[8]. Meanwhile, initiatives like BigScience's BLOOM offer models for download with extensive transparency regarding their training data and methods[9].

Despite these varying practices, all of these organisations label their models as "open-source" or "open" to emphasise their commitment to distributing the benefits of their work across the digital economy. However, some of these entities impose restrictions on the use, reproduction, or modification of their models, actions that conflict with the core principles of open-source software. This practice, known as "open washing", involves companies branding their models as "open-source" or "open" as a form of misleading virtue signalling. As Widder et. Al. explain, the term "open-source" is often used more as an aspirational or marketing tool than a technical descriptor[10]. According to Liesenfeld and Dingemanse, open washing undermines public understanding of AI, diverts funding from genuinely open projects, and fosters a culture of openness that falls short of true transparency[11]. At its heart, open washing entails suggesting that a model is open enough for its associated benefits – often related to legal exemptions – to be realised, while in reality, the model remains substantially "closed".

The complexity of defining openness in Generative AI contributes to this issue, and creates uncertainty. Unlike traditional software, where what constitutes "source code" is straightforward, Generative AI's "source code" can refer to various components, which may be published together or separately[12]. For developers to reuse or retrain a Generative AI model, additional elements like its weights and training data are essential. These nuances have led the Open Source Initiative to create a specific definition for open source AI, the first complete version of which was published in October 2024[13]. However, for now, the extent to which the new open source AI definition will reduce open washing is unclear.

Highlighting the benefits of open-source Generative AI – without specifically defining what "open source" means – is a tactic often used by companies that try to limit efforts to regulate AI. They may seek to discourage the development of AI legislation, or to secure exemptions from legislation for "open" models or Generative AI in general. For example, Meta referenced the need to protect open-source Generative AI in its opposition to the regulation of AI in both the

[7] Isaac, M. (2024). "How AI made Mark Zuckerberg popular again in Silicon Valley." *Seattle Times.* https://www.seattletimes.com/business/how-ai-made-mark-zuckerberg-popular-again-in-silicon-valley/

[8] Mensch, A. (2023). Comment on Mistral 7B v0.1 Discussion. *Hugging Face.* https://web.archive.org/web/20231221193931/https:/huggingface.co/mistralai/Mistral-7B-v0.1/discussions/8

[9] BLOOM Model Card. *Hugging Face.* https://huggingface.co/bigscience/bloom

[10] Widder, D.G., West, S. & Whittaker, M. (2023). "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." *SSRN.* http://dx.doi.org/10.2139/ssrn.4543807: p1.

[11] Liesenfeld, A. & Dingemanse (2024). "Rethinking open source generative AI: open-washing and the EU AI Act." *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.* https://dl.acm.org/doi/10.1145/3630106.3659005: p1776.

[12] Ibid: p1782.

[13] The Open Source AI Definition – 1.0. *Open Source Initiative.* https://opensource.org/ai/open-source-ai-definition

EU[14] and the US[15]. Companies such as Google[16] and Microsoft[17], two companies that do not produce "open" Generative AI models, have similarly called for exemptions for open-source models, referencing the need to protect innovation. While not all these companies have been directly accused of engaging in "open washing", this practice complicates regulatory efforts, ultimately resulting in equal amounts of fear, uncertainty and doubt.

The messaging from some of these organisations focuses almost exclusively on the benefits of openness. However, despite all its positive characteristics, openness can significantly facilitate the spread of risks associated with Generative AI. Increased transparency means developers lose control over their models, potentially allowing malicious actors to bypass safeguards and exploit the models for harmful purposes. Making Generative AI models available for download raises the risk of misuse, including the creation of illegal content, disinformation and online scams. The issue of open washing and the risks tied to the openness of Generative AI models highlight the need for a nuanced approach to openness that goes beyond questions surrounding regulatory exemptions and definitions of what "truly open" means. This paper suggests policy options that balance the risks and benefits of Generative AI by supporting a level of openness that fosters the democratic governance of Generative AI. This approach aims to maximize openness to external researchers working to enhance model safety, while restricting access to malicious actors. However, as section 3 describes, openness is a double-edged sword, whereby different degrees of openness can affect both malicious actors' ability to misuse a model, and external researchers' ability to evaluate a model and work to make it safer.

This paper's main contribution is to build on current work on the risks associated with openness by proposing a policy framework to complement overall approaches to AI risks. It highlights openness as one of the key risks that policymakers should consider in the context of AI legislation, particularly in the context of Generative AI. Although the paper focuses on Generative AI, its recommendations are applicable to any upcoming high-impact digital technologies that could be made openly available in the future.

[14] Zuckerberg, M. & Ek, D. (2024). "Mark Zuckerberg and Daniel Ek on Why Europe Should Embrace Open-Source AI: It Risks Falling Behind Because of Incoherent and Complex Regulation, Say the Two Tech CEOs." *Spotify.* https://newsroom.spotify.com/2024-08-23/mark-zuckerberg-and-daniel-ek-on-why-europe-should-embrace-open-source-ai-it-risks-falling-behind-because-of-incoherent-and-complex-regulation-say-the-two-tech-ceos/

[15] Sherman, R. (2024). *Letter to Senator Scott Wiener.* https://www.documentcloud.org/documents/25036015-sb-1047-letter-62524

[16] Google (2024). *Feedback on the EU AI Act.* https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

[17] Microsoft (2021). "Microsoft's Response to the European Commission's Consultation on the Artificial Intelligence Act." https://blogs.microsoft.com/wp-content/uploads/prod/sites/73/2021/09/microsoft-response-to-the-european-commission-consultation-on-the-artifical-intelligence-act.pdf

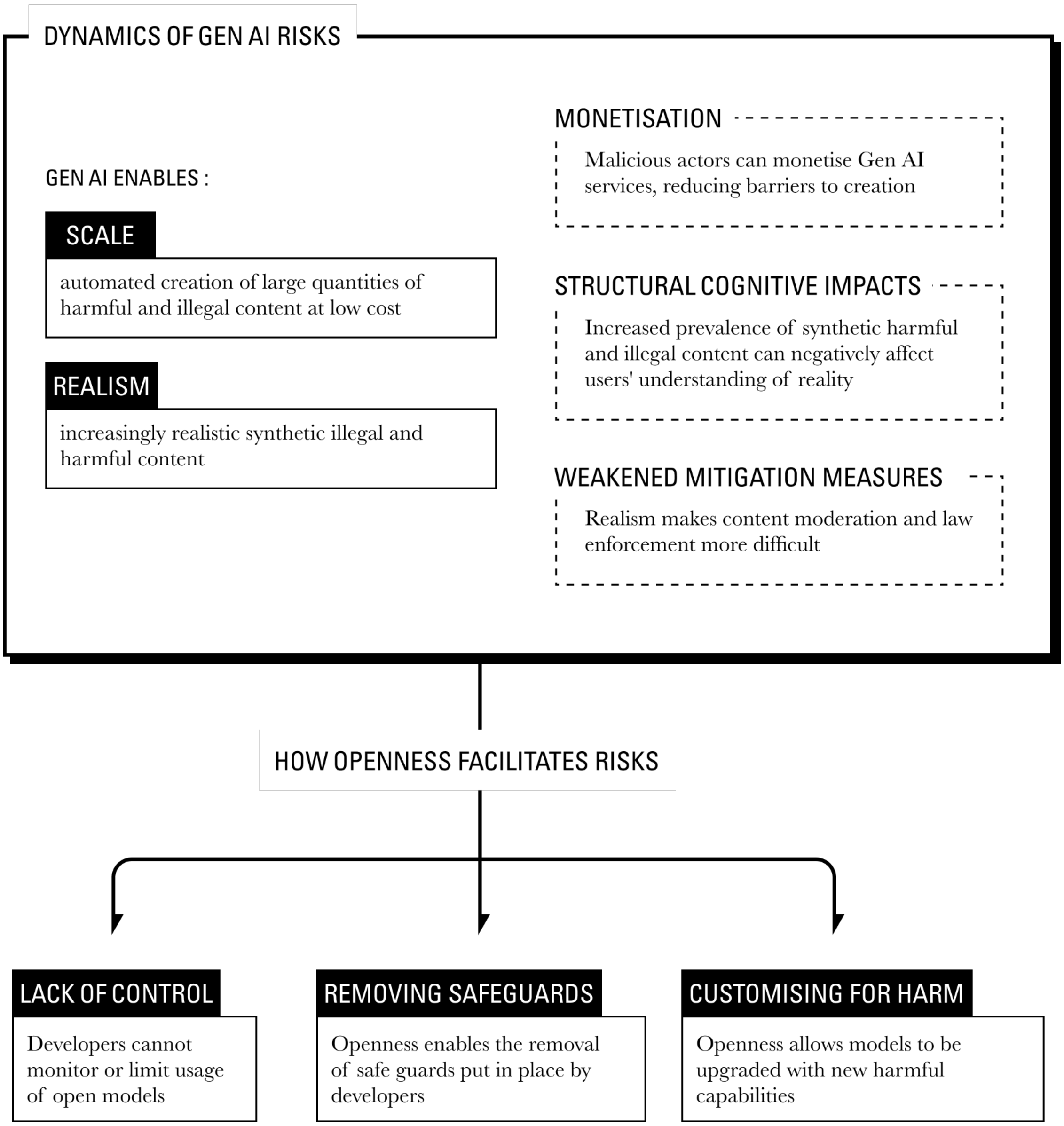
2 The Risks of Openness

Generative AI carries significant risks, which are amplified when more elements of the models are made available for download. Accordingly, this section will make the argument that, because of the risks associated with openness, a wholesale exemption from regulation for "open" Generative AI models is currently neither justifiable nor desirable.

This section will demonstrate the need for a thoughtful legislative approach to the openness of Generative AI by making the following two arguments:

- **Generative AI amplifies online safety risks**: Section 2.1 demonstrates several dynamics through which Generative AI can amplify certain online safety risks: (1) the realism of Generative AI, i.e. its ability create outputs – whether text, images, videos, or sounds – that are indistinguishable from those produced by humans, can increase the severity of certain online harms; (2) Generative AI makes the creation of harmful and illegal AI-generated content easier, increasing users' exposure to this content; and (3) increases in the scale of harmful and illegal AI-generated content can have long-term impacts on the cognitive perceptions of users. This section will explain the mechanisms through which openness propagates risks, providing examples linked to risks of misuse such as online scams, AI-generated child sexual abuse material (AI-CSAM), non-consensual intimate deepfakes (NCID) and disinformation.
- **Openness is an important facilitator of the risks associated with Generative AI**: Section 2.2 demonstrates that openness makes it easier for malicious actors to misuse Generative AI models. This is because openness (1) allows malicious actors to evade developer oversight; (2) allows those actors to remove the safeguards built into models; and (3) allows models to be upgraded and customised for harmful purposes.

These dynamics are illustrated in Figure 1.



DYNAMICS OF GEN AI RISKS
GEN AI ENABLES :
SCALE
automated creation of large quantities of harmful and illegal content at low cost
REALISM
increasingly realistic synthetic illegal and harmful content
MONETISATION
Malicious actors can monetise Gen AI services, reducing barriers to creation
STRUCTURAL COGNITIVE IMPACTS
Increased prevalence of synthetic harmful and illegal content can negatively affect users' understanding of reality
WEAKENED MITIGATION MEASURES
Realism makes content moderation and law enforcement more difficult
HOW OPENNESS FACILITATES RISKS
LACK OF CONTROL
Developers cannot monitor or limit usage of open models
REMOVING SAFEGUARDS
Openness enables the removal of safe guards put in place by developers
CUSTOMISING FOR HARM
Openness allows models to be upgraded with new harmful capabilities

FIGURE 1: HOW OPENNESS CAN EXACERBATE THE RISKS ASSOCIATED WITH GENERATIVE AI.

This paper focuses specifically on the risks of misuse, rather than structural risks associated with Generative AI, such as job displacement, or existential threats posed by AI potentially surpassing human capabilities. There are longer-term risks that are more difficult to measure, including risks related to increased government spending and the promotion of Generative AI as part of new industrial policies[18] – the impact of openness on these is less clear. The goal here is to outline some of the most important risks that are exacerbated by openness at this moment in time.

As described later in section 3, openness is a spectrum: the elements of Generative AI models (e.g. model weights, training data, etc.) and their accessibility (e.g. in a downloadable format, through an API, as a summary) can vary significantly. The most open models are released for download along with all elements, inputs and documentation, while the most closed models are not available to external users at all. The space in the middle of this spectrum is, however, more complex. For simplicity, this section will consider "open" and "closed" as a binary distinction where:

[18] AI Now Institute (2024). "AI Nationalism(s): Global Industrial Policy Approaches to AI." https://ainowinstitute.org/ai-nationalisms

- **Open models** provide sufficient elements for download so as to enable the use, scrutiny and modification of the model outside of the control of the developer. These models include Meta's Llama series, Stability AI's Stable Diffusion and BigScience's BLOOM.
- **Closed models** are those which are available only through an API, and therefore cannot be modified or used outside of the control of their developers. In this context, "closed" models include OpenAI's GPT-4o, Microsoft's Bing Chat and Google's Gemini.

The subsequent section will describe the spectrum between "fully open" and "fully closed" in greater detail.

2.1 The dynamics of Generative AI risks

This section describes the dynamics through which Generative AI can propagate or enhance online safety risks. As such, it provides a basis to analyse how openness can exacerbate these dynamics.

At its heart, Generative AI is primarily a tool that automates the creation of content. Accordingly, the main dynamics through which the technology propagates risk are **realism** (i.e. Generative AI's ability to create outputs – whether text, images, videos or sounds – that are indistinguishable from those produced by humans) and **scale** (i.e. Generative AI's ability to create large quantities of content quickly and at low cost).

Much has already been written about the risks associated with Generative AI, and to a lesser extent the role of openness. This section describes the dynamics of Generative AI risks with references to research on the following risks:

- **AI-Generated Child Sexual Abuse Material (AI-CSAM):** This refers to the use of Generative AI to create child sexual abuse material (CSAM). Under current EU law, CSAM refers to material that depicts children, any person appearing to be a child, or realistic images of children engaged in sexually explicit conduct, and depictions of their sexual organs for sexual purposes[19]. The harms associated with CSAM proliferate at different points: children are victimised when the material is created and re-victimised each time the content is viewed, shared or otherwise used. The role of Generative AI and openness in the creation of AI-CSAM is thoroughly covered in research by Internet Watch Foundation (IWF)[20] and by Thiel et al.[21].
- **Non-Consensual Intimate Deepfakes (NCIDs):** This involves using Generative AI to create deepfakes that are non-consensual intimate images (NCII), i.e. sexually explicit deepfakes created without the consent of the subject. NCIDs are a form of gender-based violence,

[19] Under a proposed revision of the Child Sexual Abuse Directive, CSAM would also include reproductions or representations of children, rather than just realistic images of children. It would also include material intended to provide guidance on how to commit child sexual abuse, exploitation or solicitation. The European Commission's goal in this revision is to include CSAM generated through Generative AI within the scope of the law. Article 2(c) Directive 2011/93/EU on combating the sexual abuse and sexual exploitation of children and child pornography, and replacing Council Framework Decision (2004/68/JHA). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52024PC0060

[20] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf

[21] Thiel, D., Stroebel, M. & Portnoff, R. (2023). "Generative ML and CSAM: Implications and Mitigations." https://purl.stanford.edu/jv206yg3793

which may harm targets' economic wellbeing (e.g. job loss), security (e.g. harassment) and personal relationships[22]. Nonetheless, in many jurisdictions, the legal status of NCIDs remains contentious[23]. The role of Generative AI in the creation of NCIDs is highlighted in research by UNESCO[24], and Frankovits and Mirsky[25].

- **Online scams:** Online fraud and scams involve cybercriminals attempting to deceive individuals into sharing sensitive information, providing access to a device, or sending money to them. These scams come in many forms. Traditional scams include phishing, where cybercriminals impersonate a trusted source to gain access to confidential information, or to infect the user's device with malware[26]. More sophisticated scams include product and service scams[27] and impersonation scams[28]. The use of Generative AI and AI in general in online scams is described in research by Hazell[29] and Mirsky et al.[30].
- **Disinformation:** In the EU, "disinformation" is informally defined as "false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm"[31]. The spread of disinformation often occurs through coordinated information influence operations, where domestic or foreign actors aim to manipulate a target audience using deceptive tactics, including the suppression of independent information sources. This can also be part of broader foreign interference efforts, which involve coercive and deceptive measures by foreign state actors or their agents to undermine the free formation and expression of individuals' political will. The role of Generative AI in disinformation campaigns is described in research by Barman et al.[32] and Bontcheva et al.[33], with research by NATO specifically focusing on the impact of openness[34].

This section's analysis of the impact of Generative AI on these risks uses a framework inspired by the AI risk framework developed by Hendrycks and Mazeika[35]. It describes how AI affects: (1) the **severity** of risks (i.e. increasing

[22] What You Need To Know About Non-Consensual Sexual Deepfakes. *Western University.* https://gbvlearningnetwork.ca/our-work/infographics/nonconsensualsexualdeepfakes/index.html

[23] Flynn, A. (2024). "Legal loopholes don't help victims of sexualised deepfakes abuse." *Monash University.* https://lens.monash.edu/@politics-society/2024/04/18/1386624/legal-loopholes-dont-help-victims-of-sexualised-deepfakes-abuse

[24] Chowdhurym R. & Lakshmi, D. (2023). "'Your opinion doesn't matter, anyway': exposing technology-facilitated gender-based violence in an era of generative AI." *UNESCO.* https://unesdoc.unesco.org/ark:/48223/pf0000387483

[25] Frankovits, G. & Mirsky, Y. (2023). "Discussion Paper: The Threat of Real Time Deepfakes." *WDC '23, July 10–14, 2023, Melbourne, VIC, Australia. arXiv.org.* https://arxiv.org/pdf/2306.02487

[26] Phishing/Spear phishing. *ENISA.* https://www.enisa.europa.eu/topics/incident-response/glossary/phishing-spear-phishing

[27] Product and service scams. *ScamWatch.* https://www.scamwatch.gov.au/types-of-scams/product-and-service-scams

[28] Federal Trade Commission (2024). "'Grandparent' Scams Get More Sophisticated." https://www.fcc.gov/grandparent-scams-get-more-sophisticated

[29] Hazell, J. (2023). "Spear Phishing With Large Language Models." *arXiv.org.* https://arxiv.org/abs/2305.06972

[30] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X, Lee, W. Elovici, Y. & Biggio, B. (2021). "The Threat of Offensive AI to Organizations." ACM Comput. Surv., 1(1). *arXiv.org.* https://arxiv.org/pdf/2106.15764

[31] Communication on the European democracy action plan. *European Commission.* https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2020%3A790%3AFIN

[32] Barman, D. & Guo, Z. (2024). "The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination." *Machine Learning with Applications*, 16. https://www.sciencedirect.com/science/article/pii/S2666827024000215

[33] Bontcheva, K. (2024). "Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities." *EDMO.* https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

[34] Haiduckyk, T., Shevtsov, A. & Bergmanis-Korāts, G. (2024). "AI in Precision Persuasion: Unveiling Tactics and Risks on Social Media." *NATO Strategic Communications Centre of Excellence.* https://stratcomcoe.org/pdfjs/?file=/publications/download/AI-In-Precision-Persuasion-DIGITAL.pdf?zoom=page-fit

[35] Hendrycks, D. & Mazeika, M. (2022). "X-Risk Analysis for AI Research." *arXiv.org.* https://arxiv.org/pdf/2206.05862

the harm), (2) users' **exposure** to risk (i.e. increasing the number of users exposed), and (3) users' **vulnerability** to risk (i.e. increasing users' susceptibility to the harm).

2.1.1 Realism

Generative AI can increase the severity of harms by making it easier for malicious actors to create realistic synthetic content. Creating realistic synthetic content previously required significant time, resources and expertise (e.g. with image editing software such as Adobe Photoshop). These barriers have effectively been reduced by Generative AI, allowing the creation of increasingly realistic, partially and fully synthetic content. This realism is an important factor in increasing the severity of certain risks associated with Generative AI.

AI-CSAM is an important example of this dynamic. Generative AI can be used to create two primary forms of AI-CSAM. First, Generative AI can be used to victimise children by "nudifying" benign pictures of fully clothed children, which might have been taken and uploaded online for legitimate reasons[36]. This creates the potential that children who have not been in direct contact with perpetrators may nonetheless become victims. Secondly, Generative AI poses a risk of re-victimising those already harmed by using AI to generate AI-CSAM based on actual depictions of abuse. In both scenarios, the realism of the AI-generated images plays a crucial role, as the content's value to perpetrators increases with its realism. Therefore, as models evolve to produce ever-more photorealistic images and videos, the risks are likely to escalate. Thiel and colleagues point out that while current models used to create AI-CSAM are generally limited to producing still images, future advancements might enable the creation of realistic, full-motion content[37].

Realism may also increase harms adjacent to **AI-CSAM** and **NCIDs** more broadly by increasing targets' vulnerability to abuse linked to sextortion and blackmail. For example, AI-CSAM created using benign pictures of children provides perpetrators the opportunity to blackmail or extort the children in question, including into sexual abuse. Similarly, the enhanced capability of Generative AI to produce realistic NCIDs heightens the risk of gender-based violence. According to Adam Dodge, founder of online safety NGO EndTAB, NCIDs are a "perfect tool for somebody seeking to exert power and control over a victim"[38]. Indeed, the US Federal Bureau of Investigation (FBI) reported that the number of sextortion schemes relying on NCIDs is increasing[39]. Whereas sextortion schemes used to only be possible with real photos and videos, Generative AI has significantly expanded malicious actors' ability to use such schemes. The realism associated with Generative AI therefore creates new avenues for perpetrators to target potential victims, making the latter more vulnerable.

Furthermore, realism plays a significant role in the risk associated with **online scams,** as Generative AI can make fraudulent messages appear authentic, thus

[36] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf: p17.

[37] Thiel, D., Stroebel, M. & Portnoff, R. (2023). "Generative ML and CSAM: Implications and Mitigations." https://purl.stanford.edu/jv206yg3793: p15.

[38] Hao, K. (2021). "Deepfake porn is ruining women's lives. Now the law may finally ban it." *MIT Technology Review.* https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

[39] Federal Bureau of Investigation (2023). "Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes." *Public Service Announcement.* https://www.ic3.gov/Media/Y2023/PSA230605

enhancing the effectiveness of social engineering[40]. For example, malicious actors can use Generative AI to ensure their communications with potential victims are more convincing. As Brundage et al. note, "[a]s AI systems grow more capable of modelling genuine human interaction, they could engage in social mimicry that could be difficult even for experts to detect"[41]. For example, Large Language Models (LLMs) can help ensure that scam and phishing emails are properly formatted and grammatically correct. Spelling mistakes have historically been a good indication that a message (e.g. from an organisation claiming to be a government agency) is fraudulent; Generative AI can help ensure that phishing emails appear legitimate[42].

This dynamic also impacts risks associated with **disinformation**. Research by Barman et al. shows that LLMs can help make disinformation content appear more real to users[43]. LLMs can be leveraged to refine and enhance disinformation content, making it subtler and more believable. These models can adapt disinformation for various formats, such as news articles or social media posts, and create convincing fake social media profiles. They can also orchestrate comprehensive dissemination strategies, including optimally timing posts, tailoring content to specific demographics, and utilising pertinent hashtags. Most significantly, when used in conjunction with bots, Generative AI can automate interactions with real users' social media posts and comments, thereby amplifying the spread and apparent legitimacy of disinformation narratives.

2.1.1.1 Weakened mitigation measures

The realism of Generative AI content may also have an impact on existing harm mitigation measures by making it more difficult for content moderators and law enforcement to distinguish between real and synthetic content.

For example, **AI-CSAM** is increasingly difficult to distinguish from real CSAM, placing a greater strain on content moderators and law enforcement. The IWF observed that while most AI-CSAM in early 2023 displayed obvious signs of artificiality, such as cartoon-like images or blurry backgrounds, by the end of the year, much of it had become life-like and challenging to distinguish from real CSAM[44]. This complexity complicates the task of law enforcement in differentiating between actual victims, those represented in AI-CSAM created from benign images, and entirely virtual representations. It also gives rise to a "liar's dividend", where perpetrators might claim that the CSAM in their possession is AI-generated.[45].

The rise of AI-generated **disinformation** similarly places a greater burden on content moderators and fact-checkers, who must now contend with both the

[40] Social engineering is defined as "using deception to convince a target to reveal information or perform certain actions for illegitimate reasons." What is social engineering? *ENISA.* https://www.enisa.europa.eu/topics/incident-response/glossary/what-is-social-engineering

[41] Brundage, M. et al. (2018). "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation." *arXiv.org.* https://arxiv.org/pdf/1802.07228.pdf

[42] Hern, A. & Milmo, D. (2023). "AI chatbots making it harder to spot phishing emails, say experts." *The Guardian.* https://www.theguardian.com/technology/2023/mar/29/ai-chatbots-making-it-harder-to-spot-phishing-emails-say-experts

[43] Barman, D. & Guo, Z. (2024). "The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination." *Machine Learning with Applications*, 16. https://www.sciencedirect.com/science/article/pii/S2666827024000215

[44] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf.

[45] Kapoor, S. et al. (2024). "On the Societal Impact of Open Foundation Models". *arXiv.org*. https://arxiv.org/pdf/2403.07918

sheer volume of disinformation and the challenge of identifying AI-generated content.

2.1.2 Scale

By making the creation of synthetic content quicker, easier and less costly, Generative AI increases the amount of this content online, increasing users' exposure to it. Starting from late 2022, Generative AI therefore became a significant growth factor for certain online safety risks.

For example, Generative AI increases the likelihood that children will become victims of **AI-CSAM**, by making it easier for malicious actors to create AI-CSAM using benign pictures, and in particular to create AI-CSAM based on existing CSAM. For example, perpetrators can download and use an open model, and fine-tune it with real CSAM content. Once set-up, Thiel et al. note that a model can generate an image in 30 seconds to 10 minutes on a consumer-grade Central Processing Unit (CPU), and in a few seconds on a high-end Graphics Processing Unit (GPU)[46]. As a result, the amount of AI-CSAM being created has grown significantly. In an internal study undertaken in the first half of 2023, Thorn found that less than 1% of all the CSAM files they detected were photorealistic AI-CSAM, although this proportion has "increased consistently since August 2022"[47]. The IWF similarly found that AI-CSAM comprised a small but growing proportion of all CSAM online[48]. They also found one forum with over 100 posts claiming to share models fine-tuned for AI-CSAM generation, all of which could potentially be used to create AI-CSAM on users' local devices, allowing them to evade detection. In one exemplary case, police in the US arrested a man who had "hundreds – if not thousands – of [AI-generated] images depicting nude or semi-clothed prepubescent minors" in his possession[49]. Like Thorn, the IWF therefore note AI-CSAM's great potential for growth.

This dynamic is similar in the case of **NCIDs**. According to Security Hero's State of Deepfakes report, the number of deepfake videos grew 550% between 2019 and 2023, 98% of which was deepfake pornography, with 99% of targets being women[50]. According to Security Hero, "It now takes less than 25 minutes and costs $0 to create a 60-second deepfake pornographic video of anyone using just one clear face image." Although NCIDs are commonly used to create nudified pictures of celebrities[51], Generative AI increases the potential that anyone may become a target of abuse. Indeed, a US-based survey by Thorn found that one in ten children report having witnessed their peers generating NCIDs of classmates[52].

[46] Thiel, D., Stroebel, M. & Portnoff, R. (2023). "Generative ML and CSAM: Implications and Mitigations." https://purl.stanford.edu/jv206yg3793: p4.

[47] Ibid: p.2-3.

[48] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf: p27.

[49] Del Valle, G. (2024). "Wisconsin man arrested for allegedly creating AI-generated child sexual abuse material / Prosecutors say Steven Anderegg made the images with Stable Diffusion and distributed them on Instagram and Telegram." *The Verge.* https://www.theverge.com/2024/5/21/24161965/ai-csam-instagram-stable-diffusion-arrest

[50] 2023 State of Deepfakes. *Security Hero.* https://www.securityhero.io/state-of-deepfakes/

[51] Maiberg, E. (2024). "AI Images in Google Search Results Have Opened a Portal to Hell." *404 Media.* https://www.404media.co/google-image-search-ai-results-have-opened-a-portal-to-hell/; Goujard, C. (2024). "Taylor Swift deepfakes nudge EU to get real about AI." *Politico.* https://www.politico.eu/article/europe-eye-fix-taylor-swift-nude-deepfake/

[52] Thorn (2024). "Youth Perspectives on Online Safety, 2023." https://info.thorn.org/hubfs/Research/Thorn_23_YouthMonitoring_Report.pdf

Additionally, Generative AI might increase individuals' exposure to **disinformation** by reducing the costs associated with generating persuasive disinformation in all its forms[53]. In the context of text-based disinformation, Musser et al. demonstrate that LLMs only need to produce usable output 25% of the time to result in cost savings for disinformation producers[54]. The cost savings compound with the volume of content generated – such as social media posts – allowing disinformation producers to achieve substantial reductions in costs. For example, with an LLM that produces usable output 75% of the time, the savings could exceed $3 million over a campaign involving 10 million social media posts, cutting costs by about 67% per post. While the cost savings for creating image, video, and voice-based disinformation are less documented, they are likely to be even more significant due to the complexity of generating synthetic visual and audio content without Generative AI.

Many individual cases exist of Generative AI being used to create disinformation content that has been distributed on all major social media platforms[55]. For example, these include AI-generated images about the conflict in Gaza[56], the 2024 French election[57], AI-generated conversations by politicians[58], and deepfakes of female politicians in bikinis[59]. However, it remains unclear how much of this content was created using open models versus closed ones. The proportion of AI-generated disinformation compared to other types of disinformation also remains uncertain.

Automating the creation of **online scams**, such as phishing emails, scam ads and fake shopping sites, can significantly reduce the labour involved, thereby increasing the number of potential targets that malicious actors can reach. Hazell notes that Anthropic's Claude 2 can generate a batch of 1,000 phishing emails in under two hours, at a cost of only $10[60]. Additionally, Generative AI-powered bots are a major factor in the rise of ad fraud, where inauthentic traffic is used to fraudulently generate advertising revenues[61].

For traditional forms of scam, UK-based cybersecurity firm Darktrace reported a 135% increase in novel forms of social engineering attacks between January and February 2023, attributing this surge to the adoption of LLMs such as ChatGPT[62]. Darktrace also noted a shift in phishing tactics, with emails asking targets to click on a link being replaced by more complex forms of phishing. In the realm of voice-cloning scams, McAfee's research involving 7,054

[53] Goldstein, J. A. et al. (2023). "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations." *Arxiv.org.* https://arxiv.org/abs/2301.04246

[54] Musser, M. (2023). "A Cost Analysis of Generative Language Models and Influence Operations." *Arxiv.org.* https://arxiv.org/abs/2308.03740

[55] Bontcheva, K. (2024). "Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities." *EDMO.* https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

[56] Klepper, D. (2023). "Fake babies, real horror: Deepfakes from the Gaza war increase fears about AI's power to mislead." *AP.* https://apnews.com/article/artificial-intelligence-hamas-israel-misinformation-ai-gaza-a1bb303b637ffbbb9cbc3aa1e000db47

[57] AI Forensics (2024). "Artificial Elections: Exposing the Use of Generative AI Imagery in the Political Campaigns of the 2024 French Elections." https://aiforensics.org/work/french-elections-2024

[58] EDMO (2024). "Prebunking AI-generated disinformation ahead of EU elections." https://edmo.eu/publications/prebunking-ai-generated-disinformation-ahead-of-eu-elections/

[59] Swenson, A. & Chan, K. (2024). "Election disinformation takes a big leap with AI being used to deceive worldwide." *AP.* https://apnews.com/article/artificial-intelligence-elections-disinformation-chatgpt-bc283e7426402f0b4baa7df280a4c3fd

[60] Hazell, J. (2023). "Spear Phishing with Large Language Models." *arXiv.org.* https://arxiv.org/abs/2305.06972: p.3

[61] Swant, M. (2024). "Updated: DoubleVerify report, Ad fraud schemes using generative AI will increase in scale, sophistication." *Digiday.* https://digiday.com/media-buying/doubleverify-report-ad-fraud-schemes-using-generative-ai-will-increase-in-scale-sophistication/

[62] Darktrace (2023). "Cyber Security Threats - Email Compromise With Generative AI." https://darktrace.com/blog/tackling-the-soft-underbelly-of-cyber-security-email-compromise

respondents across seven countries revealed that nearly a quarter had either experienced an AI-powered voice scam or knew someone who had[63]. The study also found that 77% of AI voice scam victims had lost money, pointing to the effectiveness of such scams.

2.1.2.1 Monetisation

Using Generative AI models and systems directly can be technically challenging, particularly when it comes to the generation of realistic images. Short of downloading models and fine-tuning them, users may also rely on services (e.g. through websites or decentralised channels) where more technically savvy users may create this content for them.

The service of "image-generation for hire" is particularly important in increasing the scale of **AI-CSAM** and **NCIDs**[64]. Several end-to-end encrypted chats and peer-to-peer networks exist where creators of AI-CSAM commercialise their services. In one example, the IWF found an AI-CSAM creator charging 5,000¥ (approx. $36) per month for access to 2,000 images and the ability to request ten new images on a monthly basis. Another AI-CSAM creator identified by IWF was able to generate $316.50 per month from subscriptions revenues generated from 70 users. When it comes to NCIDs, according to UNESCO[65], Generative AI has made the outsourcing of harassment easier by developing a market of "harassment for hire" services.

2.1.2.2 Structural cognitive impacts

The increased prevalence of Generative AI content may also come with structural cognitive impacts by generally affecting users' perceptions through exposure.

For example, the increased prevalence of **AI-CSAM** may also increase children's exposure to sexual abuse in general. Research suggests that CSAM users' consumption of CSAM can increase their propensity to engage in sexual acts against a child[66]. Because Generative AI may increase the overall quantity of AI-CSAM, it may therefore increase the quantity of CSAM users, and the quantity of those that try to contact children. On this point, safety-tech company Thorn argue that the "growing frequency of [AI-CSAM] generates more demand, desensitising society to the sexualization of children and growing the appetite for CSAM."

Similarly, Generative AI may increase individuals' vulnerability to **disinformation** by increasing their susceptibility to believe disinformation narratives. A study by Brookings illustrates how the spread of disinformation and the use of bots to amplify such content in a way that appears organic can distort public perception of consensus on critical issues like immigration[67]. When individuals engage with bots that convincingly propagate disinformation,

[63] McAfee (2023). "Artificial Intelligence Voice Scams on the Rise with 1 in 4 Adults Impacted." https://www.mcafee.com/ko-kr/consumer-corporate/newsroom/press-releases/press-release.html?news_id=5aa0d525-c1ae-4b1e-a1b7-dc499410c5a1&langid=48

[64] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf: p19-20.

[65] Chowdhurym R. & Lakshmi, D. (2023). "'Your opinion doesn't matter, anyway': exposing technology-facilitated gender-based violence in an era of generative AI." *UNESCO.* https://unesdoc.unesco.org/ark:/48223/pf0000387483

[66] ReDirection Project. *Protect Children Finland.* https://www.suojellaanlapsia.fi/redirection

[67] Wirtschafter, V. (2024). "The impact of generative AI in a global election year." *Brookings.* https://www.brookings.edu/articles/the-impact-of-generative-ai-in-a-global-election-year/

they become more susceptible to believing false narratives. Additionally, the growing presence of AI-generated disinformation may erode overall trust in online information, as people become increasingly sceptical of both legitimate and false content. The pervasiveness of AI-driven disinformation, combined with agents that reinforce these false narratives, could contribute to a damaging cycle of distrust in the digital information landscape.

2.2 How openness can facilitate Generative AI's risks

The openness of Generative AI models can amplify risks for several reasons: (1) developers lose control over the use of their models; (2) safeguards implemented by developers can be removed; and (3) models can be fine-tuned for harmful purposes.

2.2.1 Lack of control by developers

When models are made available through an API or gated access, developers maintain some degree of control over the model. They may monitor usage, and restrict access in response to misuse. However, when models are made openly available for download, developers lose this control, becoming unable to monitor the use of the model and to prevent its misuse. They cannot block problematic users, identify and respond to harmful usage, nor adapt the model in response to safeguard workarounds found by users ("jailbreaks"). Moreover, ensuring that improvements to open models are implemented downstream is challenging, which perpetuates model flaws and safety issues. For example, developers would struggle to introduce fixes to various jailbreaks identified by users[68]. The result of this lack of control is that malicious actors are able to create illegal and harmful content at scale with impunity.

Openly releasing a model is, at this stage, irreversible. Problematic models that have been openly released will remain available for download, even after developers have released new iterations. These older versions will continue to be accessible on model hosting platforms or, if they have been taken down, on the dark web.

Enforcing licenses for open models is difficult. As Seger et al. note, "license breaches are difficult to track and enforce when models are freely and publicly available for download" and such breaches "will also not be of great concern for malicious actors intending to cause significant harm"[69]. Licenses that restrict use cases such as Responsible AI Licenses (RAIL)[70] are therefore unlikely to prevent misuse. For example, the Responsible Use Guide[71] that was published along with the release of Llama 2 has been ignored by the creators of models

[68]ChatGPT "DAN" (and other "Jailbreaks"). *GitHub*; Grimm, D. (2024). "'Godmode' GPT-4o jailbreak released by hacker — powerful exploit was quickly banned." *Tom's Hardware.* https://www.tomshardware.com/tech-industry/artificial-intelligence/godmode-gpt-4o-jailbreak-released-by-hacker-powerful-exploit-was-quickly-banned

[69] Seger, E. et al. (2023). "Open-Sourcing Highly Capable Foundation Models." *Centre for the Governance of AI.* https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[70] Responsible AI Licenses. https://www.licenses.ai/

[71] Responsible Use Guide. *Meta Llama.* https://ai.meta.com/static-resource/responsible-use-guide/

such as Llama 2 Uncensored – a version of Llama 2 with the safeguards implemented by Meta removed[72].

Lack of oversight by developers is a particularly important factor for risks that are less responsive to safeguards. For example, bypassing protections in closed models to create **AI-generated disinformation** is relatively straightforward through relatively unsophisticated prompt engineering. The Centre for Countering Digital Hate demonstrated that by using 40 different text prompts, platforms like Midjourney, ChatGPT Plus, DreamStudio, and Microsoft's Image Creator generated disinformation images 41% of the time[73]. As highlighted by Bommasani et al.[74], the challenge of defining what constitutes disinformation complicates the implementation of effective safeguards for textual prompts, regardless of whether models are closed or open. Accordingly, using an open model would be particularly valuable for large-scale disinformation campaigns, rather than for the creation of individual pieces of content. Users attempting to create large amounts of content (e.g. images of specific politicians, social media posts about political issues) are more likely to be flagged by the moderation filters put in place by the deployer of a Generative AI system. Being able to evade such monitoring by relying on open models would therefore be highly advantageous.

2.2.2 Removing safeguards

Openly releasing a model is generally understood to usually involve publishing components that allow subsequent modifications, such as the model architecture and model weights. This makes it easier for malicious actors to fine-tune and enhance the model for harmful purposes. Competent actors can remove safeguards against misuse, such as those that would prevent the creation of harmful or illegal content. This enables them to more easily create highly realistic content at larger scale than if they simply tried to create that content through prompt engineering on closed models.

Current research indicates that removing safeguards from open models is neither resource intensive nor overly complicated. Narayanan and Kapoor argue that paying someone to fine-tune away the safeguards of an open model is not expensive[75]. Seger et al. point out that the filters that prevent Stable Diffusion from creating harmful images can be removed by deleting a single line of code[76].

As Qi et al. highlight, part of the challenge here is that introducing safeguards typically involves "embedding safety rules within pre-trained models to restrict their harmful behaviours at inference time"[77]. In other words, safety features

[72] Harris, D. E. (2023). "How to Regulate Unsecured "Open-Source" AI: No Exemptions." *Tech Policy Press.* https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/ ; Llama2-uncensored. *Ollama.* https://ollama.com/library/llama2-uncensored

[73] Center for Countering Digital Hate (2024). "Fake image factories." https://counterhate.com/wp-content/uploads/2024/03/240304-Election-Disinfo-AI-REPORT.pdf: p6.

[74] Bommasani, R. et al. (2023). "Considerations for Governing Open Foundation Models." *Stanford University.* https://hai.stanford.edu/sites/default/files/2023-12/Governing-Open-Foundation-Models.pdf

[75] Narayanan, A. & Kapoor, S. (2023). "Model alignment protects against accidental harms, not intentional ones." *AI Snake Oil.* https://www.aisnakeoil.com/p/model-alignment-protects-against; Gade, P., Lermen, S., Rogers-Smith, C. & Ladish, J. (2023). "BadLlama: cheaply removing safety fine-tuning from Llama 2-Chat 13B." *Arxiv.org.* https://arxiv.org/abs/2311.00117

[76] Seger, E. et al. (2023). "Open-Sourcing Highly Capable Foundation Models." *Centre for the Governance of AI.* https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[77] Qi, X., Zeng, Y., Xie, T., Chen, P., Jia, R., Mittal, P. & Henderson, P. (2023). "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" *Arxiv.org.* https://arxiv.org/abs/2310.03693: p.2.

are primarily implemented towards the end of model development, once most of the model capabilities have essentially been pre-determined by its training process and training data. This is the case for safeguards implemented through widely used techniques such as Reinforcement Learning from Human Feedback (RLHF)[78] and instruction tuning[79].

The ability to remove safeguards from an open model holds significant value for creators of **AI-CSAM**. If a model is capable of creating AI-CSAM because of its training data, then filters become the primary defence against such outputs[80]. While developers may be able to enhance safeguards to make it more difficult for malicious actors to remove these filters, it remains uncertain if they can make it entirely impossible. Even if a model's training data does not explicitly include CSAM or sexual content, it may still be theoretically possible to use it to create AI-CSAM, though achieving this would be highly challenging.

2.2.3 Customising for harm

The resources necessary to train models from scratch is high and increasing[81], so customising pre-trained open models can significantly reduce the resources needed to produce models with specific capabilities. At the same time, fine-tuning can be used to introduce new harmful capabilities to a model, making it capable of creating new forms of realistic harmful or illegal content.

There is ample evidence that malicious actors are fine-tuning open models, particularly to become better at producing AI-CSAM and undertaking cyberattacks. For example, the IWF has reported that older versions of Stability AI's Stable Diffusion are still being used and fine-tuned to create AI-CSAM[82]. EleutherAI's openly released GPT-J 6B model was fine-tuned to create GPT-4chan using data collected from 4chan, including hate speech[83].

Seger et al. furthermore highlight the problem of "capability overhang", whereby a model's full capabilities may not be fully known or understood by its developers prior to release[84]. A model's unexpected or unintended capabilities "can be latent within a system only to emerge unexpectedly when elicited, for example, by clever prompt engineering or integration with other software." Even with a rigorous safety alignment programme, developers might release models capable of producing certain harms, or which have certain flaws, that they cannot prevent malicious actors from exploiting.

[78] Kaufmann, T., Weng, P., Bengs, V & Hüllermeier, E. (2023). "A Survey of Reinforcement Learning from Human Feedback." *Arxiv.org.* https://arxiv.org/abs/2312.14925

[79] Muennighoff, N. et al. (2024). "Generative Representational Instruction Tuning." *Arxiv.org.* https://arxiv.org/abs/2402.09906

[80] Thiel, D. (2023). "Investigation Finds AI Image Generation Models Trained on Child Abuse." *Stanford Cyber Policy Center.* https://cyber.fsi.stanford.edu/news/investigation-finds-ai-image-generation-models-trained-child-abuse

[81] Meyer, D. (2024). "The cost of training AI could soon become too much to bear." *Fortune.* https://fortune.com/2024/04/04/ai-training-costs-how-much-is-too-much-openai-gpt-anthropic-microsoft/

[82] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf

[83] Gpt-4chan. *Hugging Face.* https://huggingface.co/ykilcher/gpt-4chan

[84] Seger, E. et al. (2023). "Open-Sourcing Highly Capable Foundation Models." *Centre for the Governance of AI.* https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models: p.18

The pool of talent that is able to modify Generative AI models small but growing[85]. Currently, any individual who has taken a graduate-level machine learning course would be able to fine-tune an open model. Running a pre-trained model requires a minimal amount of compute, and fine-tuning techniques such as Low-Rank Adaptations (LoRA), whereby a pre-trained model is trained on new data, are relatively lightweight and accessible[86].

With safeguards removed, fine-tuning open models can expand a model's ability to create different kinds of illegal content. Thiel et al. provide a list of different techniques involving the fine-tuning of open models to create **AI-CSAM**, which may also be used in the context of NCIDs[87]. The ease-of-use of these techniques varies, with some likely being out of reach for the average user. As the IWF notes, several guides nonetheless exist on the dark web with instructions on how to fine-tune models to create AI-CSAM using personal datasets, and it's likely that similar guides exist for the creation of NCIDs[88]. They also note that websites on the open web providing AI-pornography generation as a service could be abused to create AI-CSAM, despite the existence of guidelines and safeguards against the practice[89]. The IWF furthermore highlights that communities exist that collaborate to create AI-CSAM, reducing the overall burden associated with AI-CSAM creation[90]. In the context of NCIDs, Widder et al. similarly note that fine-tuning open models is not only valuable for malicious actors, but that communities exist which are using fine-tuning to create increasingly realistic NCIDs[91]. The researchers found that contributors "actively promote their contributions to gain status, proof of technical skill".

Malicious actors have benefited from being able to fine-tune open models for **online scams**. For example, FraudGPT, based on OpenAI's ChatGPT, does not include the types of safeguards that would prevent inappropriate requests. Prompted to draft a phishing email, FraudGPT will go as far as to suggest where a malicious link could be best placed[92], and it can also be used to create scam webpages[93].

[85] Seger, E. et al. (2023). "Open-Sourcing Highly Capable Foundation Models." *Centre for the Governance of AI.* https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[86] Hu, J. E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y, Wang, S., Wang, L., & Chen, W. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." *Arxiv.org.* https://arxiv.org/abs/2106.09685 ; LoRA. *Hugging Face.* https://huggingface.co/docs/diffusers/main/en/training/lora

[87] Thiel, D., Stroebel, M. & Portnoff, R. (2023). "Generative ML and CSAM: Implications and Mitigations." https://purl.stanford.edu/jv206yg3793: p4-6.

[88] Internet Watch Foundation (2023). "How AI is being abused to create child sexual abuse imagery." https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf: p.36

[89] Ibid: p.19-20.

[90] Ibid.

[91] Widder, D.G., Nafus, D., Dabbish, L. & Herbsleb, J. (2022). "Limits and Possibilities for 'Ethical AI' in Open Source: A Study of Deepfakes." *FAccT '22, June 21–24, 2022, Seoul, Republic of Korea.* https://davidwidder.me/files/widder-ossdeepfakes-facct22.pdf

[92] Amos, Z. (2023). "What is FraudGPT?" *Hackernoon.* https://hackernoon.com/what-is-fraudgpt

[93] Burns, E. (2023). "FraudGPT: The Latest Development in Malicious Generative AI." *Abnormal Security.* https://abnormalsecurity.com/blog/fraudgpt-malicious-generative-ai

3 The Impact of Degrees of Openness on Safety Risks

The previous section revealed how openness heightens the online safety risks associated with Generative AI in general. However, this analysis framed openness in binary terms, contrasting "open models" with "closed models." Openness has, however, emerged as a spectrum. Researchers Liesenfeld and Dingemanse underscore that the openness of Generative AI models is both "composite" and "gradient", in that it both consists of multiple elements and exists in degrees[94]. In line with this framework, this section will first identify the various elements of Generative AI models that can be made open. Following this, we will explore the different ways these elements can be made available and the extent to which this openness facilitates access for both external researchers and malicious actors. As such it has **a double-edged impact on safety**. On the one hand, it enhances the ability of external researchers to scrutinise these models, enabling them to identify risks and improve mitigations. On the other hand, as the previous section showed, it increases the potential for malicious actors to misuse the models in question. This dynamic represents the conflicting effects of openness on the safety of Generative AI models.

Rather than drafting regulatory exemptions, this section argues that the optimal balance lies in promoting openness to enhance external research access while simultaneously restricting the total number of users who can access the model.

3.1 Openness of Generative AI models: elements and degrees

3.1.1 Elements

Generative AI models are complex and made up of many different parts. Significant efforts have been made to catalogue the elements of Generative AI that developers make openly available. The most significant of these are the Stanford University Foundation Model Transparency Index (FMTI)[95] (which includes all forms of Generative AI models) and Liesenfeld and Dingemanse's

[94] Liesenfeld, A. & Dingemanse, M. (2024). "Rethinking open source generative AI: open-washing and the EU AI Act." *FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil.* https://pure.mpg.de/rest/items/item_3588217_2/component/file_3588218/content: p.1.

[95] The Foundation Model Transparency Index. *Stanford University.* https://crfm.stanford.edu/fmti/May-2024/index.html

"Opening up ChatGPT" framework (which focuses solely on LLMs)[96]. The elements of Generative AI models that can be made open can be split into the following categories:

- **Model inputs**: The ingredients and processes involved in creating the model, such as training data, training code, training methods and fine-tuning code.
- **The model itself**: The components of the model, such as model architecture, inference code and weights. This constitutes the finished product, which may then be used to produce outputs when integrated within an AI system and is typically accessed through an API.
- **Model outputs**: The model's responses to user prompts, which can include text, images, videos and voice content. These outputs can be accessed through an API, similar to OpenAI's ChatGPT, without the need for local downloads.
- **Documentation**: Written materials describing (among others) model inputs (e.g. categories of training data used), the model itself (e.g. evaluations of capabilities, descriptions of mitigation measures), model inputs (e.g. descriptions of training data), and model outputs (e.g. usage reports, usage policies and assessments of risks and mitigations).

3.1.2 Degrees

Degrees of openness refer to (1) the method through which access to model elements is provided, (2) the model elements that are shared and (3) the categories of users to which access is provided. The degrees of openness are described below and summarised in Table 1, based on the works of Irene Solaiman[97] and Zoë Brammer[98]. Under each individual degree of openness, a variety of the model elements described in section 3.1.1 can be made available. For example, even models that are made available through download can vary in terms of how open they are: although those who download a model will have access to the model itself and its outputs, they may not necessarily be given access to its documentation or model inputs.

This section describes how each degree of openness provides access to malicious actors on the one hand, and external researchers on the other. Accordingly, external researchers can be split into two categories:

- **Independent research**, undertaken on the basis of funding unrelated to the developing organisation, typically through "unstructured access" similar to that provided to general users.
- **Contracted research**, undertaken in exchange for remuneration on behalf of the developing organisation, usually for specific purposes such as security or bias research, and conducted through "structured access" methods tailored specifically for researchers.

For an overview of the different degrees of openness typically used by independent and contracted researchers, see Table 2.

[96] Opening up ChatGPT: tracking openness of instruction-tuned LLMs.*GitHub.* https://opening-up-chatgpt.github.io/

[97] Solaiman, I. (2023). "The Gradient of Generative AI Release: Methods and Considerations." *ArXiv.org.* https://arxiv.org/abs/2302.04844

[98] Brammer, Z. (2023). "How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access." *Institute for Security & Technology.* https://securityandtechnology.org/virtual-library/reports/how-does-access-impact-risk-assessing-ai-foundation-model-risk-along-a-gradient-of-access/

Level	*Method*	*Minimum elements*	*Researcher access*	*Malicious access*	*Examples*
1. Fully closed	Model elements are not shared with people outside of the developing organisation.	None.	N/A	N/A	DeepMind's Gopher[99]; Google's Imagen[100].
2. Query API access	AI system is hosted on developer server, and access is provided through an API.	Model outputs.	★	★	OpenAI's GPT-3 and GPT-4, Snapchat's My AI.
3. Modular API access	AI system is hosted on developer server, and access is provided through an API, allows controlled modification of the model	Model outputs, model inputs, the model itself, and documentation.	★★★	N/A	Allen Institute for AI's GROVER[101]
4. Gated downloadable access	AI system and model components are made available for download, with access "gated" through a registration process (or access is only allowed to external researchers).	The model itself, model outputs.	★★	(★★)	Meta's Llama 2 and Llama 3[102]; Mistral 7B[103].
5. Non-gated downloadable access	AI system and model components are made available for download without the need to register or pay a fee.	The model itself, model outputs.	★ ★	★★	Stable Diffusion 2 and 3[104], Vicuna 13B[105], Databricks' Dolly 12B[106]
6. Fully open	AI system, training data and protocols, documentation and components are made available for download without restriction.	Model inputs, the model itself, model outputs, and documentation.	★ ★ ★	★★★	BigScience's BLOOMZ[107] and Allen Institute for AI's OLMo 7B[108]

TABLE 1: THE DEGREES OF OPENNESS OF GENERATIVE AI MODELS, AND THE EXTENT TO WHICH THEY ENABLE EXTERNAL RESEARCHERS AND MALICIOUS ACTORS TO GAIN ACCESS TO THE MODEL IN QUESTION. ★ ★ ★ SIGNIFIES THAT THE MODEL IS HIGHLY USABLE AND MODIFIABLE; ★★ SIGNIFIES THAT IS HIGHLY USABLE, BUT HARDER TO MODIFY; ★ SIGNIFIES THAT IT IS MODERATELY USABLE AND IMPOSSIBLE TO MODIFY. FOR (4), (★★) IS INTENDED TO INDICATE THAT MALICIOUS ACTORS' ABILITY TO GAIN ACCESS TO THE MODEL DEPENDS ON HOW STRONGLY GATED IT IS.

[99] Gopher by DeepMind. *GPT-3 DEMO.* https://gpt3demo.com/apps/deepmind-gopher
[100] Imagen. *Google.* https://imagen.research.google/
[101] Grover. *Allen Institute for AI.* https://grover.allenai.org/
[102] Llama 3. *Meta.* https://llama.meta.com/llama3/
[103] Mistral 7B. *Hugging Face.* https://huggingface.co/mistralai/Mistral-7B-v0.1
[104] Stable Diffusion 2. *Hugging Face.* https://huggingface.co/stabilityai/stable-diffusion-2
[105] Vicuna 13B. *Hugging Face.* https://huggingface.co/lmsys/vicuna-13b-v1.3
[106] Dolly v2 12B. *Hugging Face.* https://huggingface.co/databricks/dolly-v2-12b
[107] BLOOMZ. *Hugging Face.* https://huggingface.co/bigscience/bloomz
[108] OLMo 7B. *Hugging Face.* https://huggingface.co/allenai/OLMo-7B

	Independent research	*Contracted research*
Fully closed	X	X
Query API access	✓	✓
Modular API access	X	✓
Gated downloadable	✓	✓
Non-gated downloadable	✓	X
Fully open	✓	X

TABLE 2: THE DIFFERENT DEGREES OF OPENNESS USED BY INDEPENDENT AND CONTRACTED RESEARCHERS.

3.1.2.1 Fully Closed

Fully closed models are those where all model elements are kept inaccessible to anyone outside the developer organisation. This approach is often used for models that are still undergoing further training and fine-tuning before their public release. Additionally, models intended solely for internal purposes, such as for enhancing productivity in specific processes, remain in this stage. Developers of fully closed models may choose not to disclose their existence, even after training is complete. In this case, neither **external researchers** nor **malicious actors** can gain access to the model.

3.1.2.2 Query API access

Query API access allows users to interact with a model through an API, a software interface that connects users remotely to the model, which remains hosted on the developer's servers. Users' interactions with the model are limited to "querying" (i.e. sending prompts and receiving responses), meaning the only element of the model shared with users through query API access is its outputs. This form of access limits users to viewing the model's outputs only and does not provide any documentation that helps users understand them. For example, commercial models such as ChatGPT are accessible through a dedicated platform[109], while other models are integrated within other services, like Snap Inc.'s My AI, which relies on ChatGPT and can be found within the Snapchat app[110].

[109] ChatGPT. *OpenAI.* https://chat.openai.com/

[110] What is My AI on Snapchat and how do I use it? *Snapchat Support.* https://help.snapchat.com/hc/en-us/articles/13266788358932-What-is-My-AI-on-Snapchat-and-how-do-I-use-it; OpenAI (2024). "What is My AI on Snapchat and how do I use it?" https://openai.com/index/introducing-chatgpt-and-whisper-apis/

Query API access provides limited opportunities for **malicious actors**. Developers maintain control by overseeing the interaction between users and the model, allowing them to implement filters against harmful inputs and adjust these filters over time in response to new threats, including in response to jailbreaks discovered by users. For instance, filters can prevent the creation of illegal content, such as AI-CSAM or deepfakes. Query API access also allows developers to monitor usage patterns, which allows them to identify problematic behaviour. Additionally, developers may implement rate limits, which would in practice limit the usefulness of their models to malicious actors who require large quantities of outputs from the model, such as in the context of persistent and long-term cyberattacks[111]. Rate limits also limit users' ability to reverse engineer a model by scraping large quantities of outputs[112].

Query API access allows **external researchers** to have access to Generative AI models to the same extent as malicious actors, allowing them to test models for vulnerabilities and risks. Independent researchers will typically be subject to the same rate limits and content moderation restrictions as malicious actors, which may restrict the scope of their research. Contracted researchers, meanwhile, may have rate limits lifted by developers, and be subject to a form of safe harbour from content moderation (this is discussed further in section 4.3).

3.1.2.3 Modular API access

Modular API access, sometimes referred to as "researcher API access", is similar to query API access in that it involves making a model available through an API, and potentially allowing users to modify the model. This access is typically reserved for **external researchers**, especially those contracted by the developing organisation, and includes fewer restrictions compared to query API access. For instance, restrictions applicable to query API access such as rate limits will typically not apply. Researchers may be able to fine-tune the model based on their findings, though developers retain oversight to monitor and prevent harmful changes. Modular API access may be revoked if unforeseen risks arise.

Modular API access nevertheless comes with limitations primarily linked with developers remaining in control of the extent to which a model can be examined and modified, as discussed in further detail in section 4.3.1.

Malicious actors are typically unable to access Generative AI models through modular API access, as this form of access is typically reserved for vetted researchers. However, the risk of reverse engineering increases with this deeper level of access, particularly if additional elements like training methodologies or data are provided[113]. Therefore, modular API access is usually granted only after a thorough evaluation and vetting process by the developing organisation, making it usually not accessible to independent researchers.

[111] Rate limits. *OpenAI.* https://platform.openai.com/docs/guides/rate-limits

[112] Asnani, V., Yin, X., Hassner, T. & Liu, X. (2023). "Reverse Engineering of Generative Models: Inferring Model Hyperparameters From Generated Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45(12).https://ieeexplore.ieee.org/abstract/document/10202583

[113] Brammer, Z. (2023). "How Does Access Impact Risk? Assessing AI Foundation Model Risk Along a Gradient of Access." *Institute for Security & Technology.* https://securityandtechnology.org/virtual-library/reports/how-does-access-impact-risk-assessing-ai-foundation-model-risk-along-a-gradient-of-access/: p19.

3.1.2.4 Gated downloadable access

Gated downloadable access allows users to download the model to their local device, where they can run and modify it. Access can be "gated" to different extents: access to Meta's Llama 3 is gated by a simple registration procedure[114], which facilitates access to all categories of users. Other models may be gated more strictly, as part of pre-release evaluations and testing, in which case only external researchers would be able to gain access.

When downloadable access is provided exclusively to **external researchers**, whether contracted or independent, it allows them to undertake even more robust research than under modular API access[115]. However, gated downloadable access on its own does not imply full access to model inputs (e.g. training data and methodology), only to the model itself. That means that while it is possible for users downloading the model to build upon it, replicating it is difficult. In practice, gated downloadable access is difficult to revoke (e.g. after a researcher's contract has ended), as users are able to copy the model.

Malicious actors may gain access to models through gated downloadable access in two ways: (1) if access is gated through a simple registration process and (2) if the model is provided through a stricter gating process, but nonetheless leaked, copied, or reverse engineered by the individuals to which downloadable access was provided. In the first instance, simple terms and conditions and/or a license are unlikely to deter actors that wish to use the model to nefarious ends. In the second instance, malicious actors' ability to gain access will depend on the contractual protections that the developer has put in place to prevent the model from being leaked. Here, the fear of legal repercussions may be sufficient to deter model leaks.

With the model downloaded on their local device, malicious actors' ability to misuse the model increases exponentially, and developers' ability to prevent misuse effectively disappears. At this point, developers stop being able to prevent malicious actors from using the model, removing its safeguards and fine-tuning it.

3.1.2.5 Non-gated downloadable access

Non-gated downloadable access refers to models that can be downloaded without any restrictions, payment, or terms and conditions. This type of access makes it impossible for developers to track who downloads the model. Non-gated downloadable access is in practice impossible to revoke, either technically or legally, given the absence of technical infrastructure or legal mechanism to do so. Non-gated downloadable access does not necessarily entail providing users with all model inputs and documentation, which may limit users' ability to modify the model.

Non-gated downloadable access allows all varieties of **external researchers** to download the model in question without any restrictions, as well as **malicious actors.**

[114] Llama 3. *Meta.* https://llama.meta.com/llama3/

[115] Solaiman, I. (2023). "The Gradient of Generative AI Release: Methods and Considerations." *ArXiv.org.* https://arxiv.org/abs/2302.04844: p6.

3.1.2.6 Fully open

Fully open models are those where the model itself, its inputs and documentation are made available to users without any restrictions, payment or terms and conditions. This level of openness allows for the maximum understanding of the model and the ability to replicate or build upon it. However, few models are truly "fully open" since developers often withhold key elements like training data. As with other forms of downloadable access, revoking the openness of "fully open" models is impossible.

External researchers' ability to understand a model increases significantly once it is made fully open. It allows them to fully understand and trace the origins of model outputs while they test the model and its safeguards.

Malicious actors' ability to modify a model similarly increases as more of its elements and documentation are made available for download. Sufficiently capable actors may even go as far as being able to retrain the model, in view of making it more effective at achieving certain ends. However, because of the large amount of resources and skills required to train Generative AI models, few malicious actors beyond state actors would be able to retrain a Generative AI model.

4 Finding a Balance through Openness to External Researchers

Openness is crucial for shedding light on Generative AI models, allowing scrutiny beyond the confines of their creators. When models are made available for download, users can interact with them without the limitations imposed by APIs controlled by their developers, which typically restricts the extent of user interaction. However, full access to all model elements is not always necessary for achieving the safety benefits associated with openness. Section 3 suggests that there exists a "sweet spot" in the degrees of openness of Generative AI models, where developers can optimally balance the benefits and risks. This sweet spot involves opening up Generative AI models sufficiently so that external researchers may work to make the model safer, while keeping the model closed to users that may misuse it – at least until it can be considered sufficiently safe. Achieving this involves a mix of query API access, modular API access, and gated-downloadable access – collectively known as "structured access." This approach involves granting special access to researchers while restricting broader use until the model is deemed "safe".[116]

This section suggests that promoting openness to external researchers should be an integral part of legislation applicable to Generative AI. This is because:

- **External researchers can contribute to making models safer**: Section 4.1 describes the various ways in which external researchers can test and evaluate models, their strengths, weaknesses and impacts.
- **External researcher access is a means of securing the safety benefits of openness**: Section 4.2 describes how openness to external researchers can promote the democratic governance of Generative AI, help make individual models become safer, boost innovation in the development of safeguards through an open science approach, and drive the development of safety norms in the long-term.
- **Important limitations currently reduce the benefits of openness to external researchers**: Section 4.3 describes various limitations linked to the activities of developers and the resources available to researchers which currently prevent this level of openness from reaching its full potential. These limitations should be the focus of policymakers' attention.

[116] Shevlane, T. (2022). "Structured access: an emerging paradigm for safe AI deployment." *ArXiv.org.* https://arxiv.org/abs/2201.05159

4.1 A snapshot of external AI research activities

"External research" is a broad term that applies to a wide variety of multidisciplinary activities. Research into Generative AI models can involve interactions between the fields of computer science and engineering, law, social science[117], economics[118], media studies[119] and organisational studies[120] – to name a few.

In general, external research is conducted by the following groups: academics, research institutes, government agencies, industry bodies (e.g. established by industry for this particular purpose), non-profit organisations, and international organisations. As noted in section 3.1.2, it can be undertaken voluntarily or based on a contract with the developing organisation. They may be mandated by law in the form of an audit, in which case a government agency or an authorised third-party will undertake it.

In the context of Generative AI, external research can focus on the following (often overlapping) activities:

- **Model capabilities**: Evaluating the model's performance and limitations (e.g. on specific tasks). This may include benchmarks focusing on specific areas (e.g. language understanding, legal reasoning, scientific problems) to compare the model's capabilities with those of others[121]. This may include functionality audits, which evaluate model performance for specific applications[122].
- **Model controllability**: Assessing whether the model and its safeguards operate according to the developers' intentions. This includes "red teaming", where researchers simulate adversarial scenarios to test the model's defences and uncover potential vulnerabilities.[123] This type of research can help inform which model safeguards should be implemented and adapted as part of testing.
- **Model impacts:** Analysing how the model affects users and society. This involves evaluating both the model's elements (e.g. its outputs and training data) and its broader societal impacts (e.g. effects on specific populations or sectors like education[124], employment[125], law[126] or issues like privacy[127] and human rights[128]).

[117] Castelle, M. (2020). "The Social Lives of Generative Adversarial Networks." *University of Warwick Centre for Interdisciplinary Methodologies*. https://castelle.org/pdfs/Castelle%202020-The%20Social%20Lives%20of%20Generative%20Networks-Full%20Preprint-20200129.pdf

[118] UK Competition and Markets Authority (2023). "AI Foundation Models: Initial report." https://www.gov.uk/government/publications/ai-foundation-models-initial-report

[119] Sag, M. (2023). "Copyright Safety for Generative AI." *Houston Law Review 61(2).* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593

[120] Floridi, L. et al. (2022). "capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act." *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4064091

[121] McIntosh, T. R., Susnjak, T., Liu, T., Watters, P. & Halgamuge, M. (2024). "Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence." *ArXiv.org.* https://arxiv.org/pdf/2402.09880

[122] Kroll, J. A. (2018). "The fallacy of inscrutability." *The Royal Society Publishing.* https://royalsocietypublishing.org/doi/10.1098/rsta.2018.0084

[123] Wisbey, O. (2024). "AI red teaming." *TechTarget.* https://www.techtarget.com/searchenterpriseai/definition/AI-red-teaming; Lin, L. et al. (2024). "Against The Achilles' Heel: A Survey on Red Teaming for Generative Models." *ArXiv.org.* https://arxiv.org/abs/2404.00629

[124] Smolansky, A. et al. (2023). "Educator and Student Perspectives on the Impact of Generative AI on Assessments in Higher Education." *L@S '23: Proceedings of the Tenth ACM Conference on Learning @ Scale.* https://dl.acm.org/doi/abs/10.1145/3573051.3596191: p. 378-382.

- **Model governance:** Assessing the organisational and management structures and procedures of the developing organisation. This includes evaluating quality management systems, recording and documentation procedures, adherences with best practices, and roles and responsibilities.
- **Model compliance**: Assessing the model's legal compliance, focusing on areas such as privacy (e.g. does the model leak personal data?), copyright (e.g. was the model trained using copyrighted content?) and the possibility that the model may produce illegal content (e.g. CSAM).

External research may be undertaken at various stages (e.g. during pre-deployment) as part of the development lifecycle, and in interaction with the design, re-design, testing, training and fine-tuning of the model[129]. But it may also be undertaken post-deployment, as a means of verifying development claims, mitigating harms and identifying emergent risks not foreseen by the developer. Additionally, external research can be used as an ongoing governance mechanism, whereby the behaviour and performance of a model can be monitored over time.

4.2 External research as a means of securing the safety benefits of openness

Expanding the extent to which Generative AI models are made open to external researchers is crucial to democratising their governance. Currently, the development and release of AI models are predominantly controlled by the developers themselves, often without sufficient external oversight. This insular approach allows developers to unilaterally decide whether their models are "safe" for release, creating a significant conflict of interest. As noted by the Ada Lovelace Institute, this lack of external validation effectively means developers are "marking their own homework", potentially prioritising commercial benefits over genuine safety and ethical considerations [130].

Increasing developers' engagement with external researchers not only enhances transparency but also fosters a more diverse and inclusive approach to Generative AI governance. As Kapoor et al. highlight, this diversity of viewpoints is vital for mitigating the risks associated with models with larger market shares. Without external input, there is a risk that a monoculture could emerge, where vulnerabilities in one model might have far-reaching, systemic

[125] Hui, X., Reshef, O. & Zhou, L. (2023). "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market." *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4527336

[126] Kolkman, D. Bex, F. & van der Put, M. (2024). "Justitia ex machina: The impact of an AI system on legal decision-making and discretionary authority." *Big Data & Society.* https://journals.sagepub.com/doi/full/10.1177/20539517241255101

[127] Garante per la Protezione dei Dati Personali (2024). "ChatGPT: Garante privacy, notificato a OpenAI l'atto di contestazione per le violazioni alla normativa privacy." https://gpdp.it/home/docweb/-/docweb-display/docweb/9978020#english

[128] Darnton, H., Andersen, L., Hoh, J. Y. & Nigam, S. (2024). "A Human Rights Assessment of the Generative AI Value Chain." *BSR.* https://www.bsr.org/en/blog/a-human-rights-assessment-of-the-generative-ai-value-chain

[129] Kangeter, L. (2024). "A Lifecycle Approach to AI Risk Reduction: Tackling the Risk of Malicious Use Amid Implications of Openness." *Institute for Security and Technology.* https://securityandtechnology.org/virtual-library/reports/a-lifecycle-approach-to-ai-risk-reduction/

[130] Ada Lovelace Institute and other signatories (2023). "Post-Summit civil society communique." *Ada Lovelace Institute.* https://www.adalovelaceinstitute.org/news/post-summit-civil-society-communique/

consequences[131]. For example, in the event that a single model begins being integrated within the systems of a large number of companies, any associated vulnerabilities could propagate throughout the whole digital economy.

This section will illustrate how openness to external researchers (1) leverages collective expertise to identify and address safety concerns; (2) drives the popularisation of innovations in model safeguards; and (3) can aid in establishing responsible release standards, promoting a culture of due diligence around how models are made available to the public.

4.2.1 Crowdsourcing safe development

At its heart, involving external researchers in AI model development is about amplifying efforts to enhance model safety through crowdsourcing. Given the rapid evolution of AI models[132] – along with their increasing applications and risks – it's nearly impossible for internal developer teams to anticipate every potential use case and issue alone. While developers can foresee model performance to an extent, predicting how these models will be applied as they evolve and learn over time is a significant challenge. Crowdsourcing safety research helps bridge this gap, recognising that guiding model development away from risky behaviours is a complex and often imperfect task[133].

Moreover, involving external researchers ensures a broader range of perspectives, as Liang et al. note[134]. By tapping into diverse institutions, languages, cultures and disciplines, developers can critically evaluate models from multiple viewpoints, enhancing their safety across various contexts. Developers' internal safety teams would generally be too small and homogenous to adequately do so on their own[135].

This expanded scope of research helps developers more accurately define the appropriate uses for their models. It can inform decisions about licensing models for specific sectors or cultural contexts, and guides downstream users on how – or whether – to build upon these models[136]. Ultimately, this collaborative approach not only broadens the safety net but also clarifies the intended and unintended uses of AI models, paving the way for more informed and responsible deployment. The highly complex nature of Generative AI nonetheless complicates risk mitigation. Unlike traditional software, which often requires simple patches for vulnerabilities, AI models may have fundamental issues tied to their training data, algorithms and fine-tuning processes. Certain vulnerabilities and risks identified post-release could justify a costly form of retraining and/or the retraction of the model (e.g. in case it is being implemented in business clients' systems). This suggests that crowdsourcing improvements to make models safer should begin well before a model's public release.

[131] Kapoor, S. et al.(2024). "On the Societal Impact of Open Foundation Models". *arXiv.org*. https://arxiv.org/pdf/2403.07918

[132] Stanford University (2024). "The AI Index Report." https://aiindex.stanford.edu/report/

[133] Shah, R., et al. (2022). "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals." *Arxiv.org.* https://arxiv.org/abs/2210.01790

[134] Liang, P., Bommasani, R., Creel, K. & Reich, R. (2022). "The Time Is Now to Develop Community Norms for the Release of Foundation Models." *Stanford University.* https://crfm.stanford.edu/2022/05/17/community-norms.html

[135] Wiggers, K. (2024). "OpenAI's new safety committee is made up of all insiders." *TechCrunch.* https://techcrunch.com/2024/05/28/openais-new-safety-committee-is-made-up-of-all-insiders/

[136] Danish Contractor, McDuff, D. Haines, J., Lee, J., Hines, C., Hecht, B, Vincent, N. & Li, H. (2020). "Behavioral Use Licensing for Responsible AI." *Arxiv.org.* https://arxiv.org/abs/2011.03116

4.2.2 Innovation in model safeguards

In the realm of Generative AI, the principle of openness can be a catalyst for significant innovation, particularly in the development of model safeguards. By broadening researchers' access to these models, industry is capable of not only advancing the scientific understanding surrounding them but also accelerating the deployment of critical safety measures that protect users and systems alike.

Enabling a diverse group of experts to examine and contribute to Generative AI models significantly boosts the development of innovative safety solutions. External researchers bring varied perspectives and expertise, leading to more effective safeguards and accelerating the creation of best practices that can be applied across the industry.

This however does imply taking an "open science" approach to the use of external researchers, whereby the knowledge they generate is made publicly accessible[137]. Such a level of transparency is possible throughout the different degrees of openness described in section 3. Whether a model is made available through query API access or downloadable access, freely publishing (for example) the results of evaluations of model capabilities, downstream impacts and controllability can provide a significant benefit to the rest of the market.

Accordingly, openness to external researchers can be a means of growing innovation in the field of AI safety, in tandem with innovation in model capabilities. It can help promote the development and testing of new model safeguards. This can include, for example, "concept erasure" techniques, whereby models become unable to produce certain content categories[138]. It can also aid the development of watermarking techniques that are tamper-proof, and resistant to subsequent fine-tuning[139]. It can furthermore inform the development of broader safety-by-design frameworks[140], which describe measures that developers should take throughout the model development lifecycle.

4.2.3 Defining standards for responsible release

External research could play a pivotal role in setting standards for the responsible release of Generative AI models. Currently, developers are free to decide the degree of openness under which their model should be released (as described in section 3). Improving external scrutiny into Generative AI models can help developers and regulators determine the conditions under which a model can be considered safe enough to be made open to a high degree (e.g. non-gated downloadable access).

Some models are inherently safer to openly release than others. For example, a developer may have implemented sufficiently robust safeguards to meaningfully

[137] Vincent-Saez, R. & Martinez-Fuentes (2018). "Open Science now: A systematic literature review for an integrated definition." *Journal of Business Research 88,* https://www.sciencedirect.com/science/article/abs/pii/S0148296317305441: p428-436.

[138] Pham, M. et al. (2023). "Circumventing Concept Erasure Methods For Text-to-Image Generative Models." *ArXiv.org.* https://arxiv.org/abs/2308.01508

[139] Qiao, T. et al. (2023). "A novel model watermarking for protecting generative adversarial network." *Computer & Security 127.* https://www.sciencedirect.com/science/article/abs/pii/S0167404823000123

[140] Thorn & All Tech is Human (2024). "Safety by Design for Generative AI: Preventing Child Sexual Abuse." https://info.thorn.org/hubfs/thorn-safety-by-design-for-generative-AI.pdf

prevent or hinder misuse. Additionally, some models have narrower scopes or more limited capabilities, so that it would simply be impossible to misuse (e.g. a text-generator cannot create AI-CSAM). This implies that it may be possible to identify minimum safety thresholds – such as those related to capability, controllability and scope – that could be used to assess whether a model is safe for open release. Accordingly, standardising these thresholds could help developers decide how open their models should be. For example, a developer could determine that their model is only safe enough to be released through query API access, rather than downloadable access. Additionally, these standardised thresholds could help developers chart a path towards greater levels of openness by highlighting areas of improvement and describing best practices.

However, the work needed to identify minimum safety thresholds is complex and dependent on model types and the context in which they will be used. What constitutes "sufficiently safe" will depend on the type of content a model can generate (i.e. text, images, video or sound), its capabilities and its scope. For example, text-generation models highly specialised towards customer service tasks would have a relatively low threshold for responsible release (e.g. it cannot produce disinformation because it is unable to answer questions about history)[141]. Meanwhile, models that are highly capable and come closer to the definition of "general purpose technologies"[142] are likely to have much higher thresholds for responsible release (e.g. OpenAI's GPT4). It is also possible that certain highly capable models may never have strong enough safeguards to justify full openness. In this context, query API access is the only degree of openness that could constitute "responsible release".

Accordingly, given the complexity of predicting all potential risks, their mitigations, and the circumstances under which they may arise, external researchers' expertise is crucial. Their input can help developers establish and validate responsible release standards, ensuring that models minimise misuse risks before they are made publicly available. This collaborative effort can provide developers, regulators, and consumers with greater assurance that AI models are safe and responsibly released. Responsible release standards therefore constitute an important ambition that could be realised through openness to external researchers.

4.2.3.1 Responsible release standards could eventually become legally mandated

In the same way that the EU AI Act requires high-risk AI systems to meet certain safety, transparency and organisational standards prior to an AI system being placed on the market[143], openly releasing Generative AI models could also be made conditional on the model in question meeting certain standards. This would entail setting specific safety, transparency and organisational standards that models must meet before they are openly released. Such standards would define the conditions under which a model is deemed safe for higher degrees of openness. It could also stipulate that models too risky for open

[141] However, less capable models that include design flaws can still carry significant risks if widely adopted, given the increased exposure to the risks in question.

[142] Crafts, N. (2021). "Artificial intelligence as a general-purpose technology: an historical perspective." *Oxford Review of Economic Policy 37(3).* https://academic.oup.com/oxrep/article/37/3/521/6374675: p521–536.

[143] *AI Act.* Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

release should only be released at lower degrees of openness (e.g. query API access).

4.3 The limitations of openness to external researchers

There are nonetheless a number of limitations that hinder the benefits associated with openness to external researchers. These are: developer control over the scope of external research, researcher resource limitations, insufficient access to data that is essential to research, and lack of safe harbours for external research.

Accordingly, these are the main areas that should be tackled by a policy framework that seeks to balance the risks and benefits of openness.

4.3.1 Developer control over the scope of external research

The way in which external researchers gain access to Generative AI models can significantly impact the scope of their research. This access is often shaped by the commercial priorities and constraints of the organisation developing the model in question, which in turn can narrow the focus of the research. Harrington and Vermeulen highlight significant differences in how Generative AI developers approach external research[144]. For example, OpenAI has cast a wide net, engaging researchers from diverse fields such as fairness, cybersecurity and disinformation. In contrast, the external research efforts that Inflection AI has publicly disclosed have been notably narrower, focusing exclusively on biosecurity and mental health.

This disparity may stem from the inherent challenges of working with general-purpose technologies like Generative AI. Given the broad spectrum of potential applications for these models, developers may only be able to anticipate a fraction of their eventual uses. Consequently, both internal and external research efforts tend to concentrate on areas that developers deem most pressing.

Moreover, developers may deliberately limit the focus of research. As Narayanan and Kapoor suggest, developers might avoid exploring certain areas of potential misuse to sidestep revealing vulnerabilities they are ill-equipped to address[145]. This selective approach can result in models that advance developers' commercial interests without necessarily fully addressing all safety concerns.

In summary, the scope of research into Generative AI models is intricately linked to how access is granted and managed. While openness can encourage diverse and innovative research, the limitations imposed by developers and the resource constraints faced by independent researchers can significantly shape

[144] Harrington, E. & Vermeulen, M. (2024). "External researcher access to closed foundation models." *Mozilla.* https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf

[145] Narayanan, A. & Kapoor, S. (2024). "AI safety is not a model property." *AI Snake Oil.* https://www.aisnakeoil.com/p/ai-safety-is-not-a-model-property

the nature and effectiveness of safety advancements in this rapidly evolving field.

4.3.2 Researcher resource limitations

Whereas contracted researchers typically benefit from having their research funded by the developers of the models they are evaluating, independent researchers must seek funding from external institutions, such as academic institutions, governments and foundations. This means that the scope of their research can be limited by their ability to secure funding.

Resource limitations mean that high degrees of openness do not inherently ensure rigorous safety research. As Widder et al. point out, AI research is resource-intensive, and merely providing access does not guarantee that independent experts will volunteer their time and resources for thorough scrutiny[146]. Seger et al. underscore that participating at the forefront of AI research demands substantial financial, computational and data resources – assets that are often beyond the reach of individuals or smaller organisations[147]. Additionally, the talent pool able to undertake AI safety research is small, particularly in Europe[148]. This scarcity of resources and expertise further complicates the landscape, highlighting the need for more structured and inclusive approaches to advancing AI safety.

4.3.3 Insufficient information access

In addition to constraints on the scope of external research, developers often restrict the types of information and model elements available to outside researchers. While different research types (see section 4.1) require access to different model elements, insufficient access can significantly limit the conclusions researchers can draw and may even lead to misleading, inaccurate or unreproducible results. For instance, while unrestricted access to model outputs might suffice for evaluating a model's capabilities, interpretability research – essential for understanding how models arrive at their outputs – requires access to the underlying model inputs[149]. Similarly, assessments related to training data become much more challenging without direct access to that data[150].

The problem of insufficient information access has worsened as a result of a troubling trend that emerged in 2023, whereby commercial model developers have become more reluctant to provide comprehensive technical documentation along with their models[151]. This shift, driven by fears that detailed documentation could be exploited by competitors, further constrains

[146] Widder, D.G., West, S. & Whittaker, M. (2023). "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." *SSRN.* http://dx.doi.org/10.2139/ssrn.4543807

[147] Seger, E. et al. (2023). "Open-Sourcing Highly Capable Foundation Models." *Centre for the Governance of AI.* https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

[148] In a 2023 Deloitte survey, 39% of senior executives in Europe named "lack of technical talent and skills" as the leading obstacle to developing and deploying generative tools/applications. Winters, S., Horton, R. & Corduneanu, R. (2024). "Now decides next. Is Europe ready for generative AI?" *Deloitte. https://www2.deloitte.com/ro/en/pages/technology/articles/state-generative-ai-enterprise-now-decides-next.html*

[149] Bucknall, B. S. & Trager, R. F. (2023). "Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements." *University of Oxford.* https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements

[150] Casper, S. et al. (2024). "Black-Box Access is Insufficient for Rigorous AI Audits." *arXiv.org.* https://arxiv.org/pdf/2401.14446

[151] Benaich, N. (2023). "State of AI Report 2023." https://www.stateof.ai/

the independent research community and impedes efforts to thoroughly evaluate model safety and efficacy.

For contracted research, the degree of access granted is also a crucial factor. Developers often restrict the tasks researchers can perform, especially when access is provided through APIs[152]. For example, query API access or modular API access can limit researchers to specific tasks defined by their contracts, which not only reduces the likelihood of uncovering unexpected vulnerabilities but also allows developers to control what information is disclosed. Even with modular API access, developers might impose limits on the extent to which external researchers can fine-tune the model, primarily due to commercial concerns about diminishing the model's market value[153]. In contrast, access through (gated) downloadable models tends to have fewer restrictions, offering broader opportunities for research and evaluation.

4.3.4 Lack of safe harbour

A significant obstacle to independent research in Generative AI is the lack of safe harbours that protect researchers from potential legal repercussions. As noted by Longpre et al., the terms of service for many leading AI systems explicitly prohibit independent external research, with violations potentially leading to account suspensions or bans. While these restrictions are intended to prevent malicious use, they also deter legitimate research by instilling a fear of legal consequences among researchers.[154]

Furthermore, as Harrington and Vermeulen highlight, some Generative AI developers even lack a safe harbour linked to their vulnerability reporting programs[155]. While companies like OpenAI, Anthropic, Google, and Cohere all offer mechanisms for reporting vulnerabilities, only OpenAI provides explicit safe harbour protections. This lack of legal immunity means that external researchers and even well-meaning users who identify vulnerabilities might face legal consequences despite their good intentions. This gap underscores the need for clearer, more supportive frameworks to encourage and protect independent research in the rapidly evolving field of Generative AI.

The issue of safe harbours is particularly acute in the realm of illegal content, the creation of which could result in more significant legal penalties for external researchers. For example, testing whether a model is capable of producing illegal content would naturally see researchers attempt to create it. Fear of legal consequences associated with the possession of illegal content could discourage external researchers from attempting to get a model to produce it, which could ultimately result in the model being less safe.

[152] Solaiman, I. (2023). "The Gradient of Generative AI Release: Methods and Considerations." *ArXiv.org.* https://arxiv.org/abs/2302.04844: p5.

[153] Robertson, A. (2024). "You sound like a bot." *The Verge.* https://www.theverge.com/24067999/ai-bot-chatgpt-chatbot-dungeon

[154] Longpre, S. et al. (2024). "A safe harbor for AI evaluation and red teaming." *arXiv.org.* https://arxiv.org/abs/2403.04893

[155] Harrington, E. & Vermeulen, M. (2024). "External researcher access to closed foundation models." *Mozilla.* https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf

5 Current Policy Approaches to Openness

Section 4 demonstrated that there are significant benefits associated with greater openness to external researchers, but that there are also significant barriers to securing these benefits. This section will lay the groundwork for policy recommendations to tackle these limitations (section 6) by describing existing policy approaches to the openness challenge. It will describe the EU's approach, touching on the AI Act and the Digital Services Act (DSA), the UK's approach, and the US's approach.

5.1 The EU's approach

5.1.1 The AI Act

5.1.1.1 Description

The EU approach to the governance of Generative AI, and AI in general, has focused on the development of rules to curb the risks associated with AI. The culmination of this work is the AI Act[156] which will become fully applicable in 2026. Its purpose is to ensure AI is safe and trustworthy. The AI Act is also intended to promote innovation and competitiveness by harmonising EU Member States' approach to AI, and by embracing a risk-based approach that places restrictions only on the most high-risk uses of AI (rather than all AI systems).

The AI Act uses a risk-based approach with different obligations for developers and users of AI models and systems. Along with bans on certain AI practices and transparency obligations for specific uses of AI deemed of limited risk, the AI Act includes: (1) obligations for users and developers of "high-risk AI systems", (2) obligations for developers of General Purpose AI (GPAI), including Generative AI. These two sets of obligations have different implications for the openness of Generative AI models.

The obligations on "high-risk AI systems" apply to a wide variety of uses of AI, including AI used in recruitment, toys and machinery. This could include systems that rely on Generative AI, such as an application that relies on an LLM to analyse CVs, or a voice-generation model integrated within a speaking toy. The obligations in question include implementing a risk management

[156] *AI Act.* Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

system, automatic activity recording, drawing up and maintaining technical documentation, data governance requirements, human oversight requirements, and accuracy and robustness requirements.

Additionally, certain deployers of "high-risk AI systems" should conduct a "fundamental rights impact assessment". This includes deployers of "high-risk AI systems" that are public bodies, private bodies providing public services, as well as deployers of AI-based credit scoring systems, and life and health insurance risk assessment systems[157]. A fundamental rights impact assessment entails descriptions associated with the use of the AI system, the categories of individuals likely to be affected by its use, the risk of harms likely to impact them, human oversight measures, and risk mitigation measures. Once performed, the fundamental rights impact assessment must be notified to the relevant market surveillance authority.

The obligations on GPAI developers apply to AI models that display "significant generality" and which are "capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications"[158]. Accordingly, the majority of Generative AI models are likely to fall within this definition. GPAI developers are required to draw up and publish a "sufficiently detailed summary about the content used for training" of the GPAI model. They are also required to draft and maintain technical documentation for the purpose of making it available to regulators and to their business users (i.e. organisations that integrate the GPAI within their services)[159]. This requirement does not apply to models made publicly available under a "free and open licence" along with their "parameters"[160]. This exception does not apply to GPAI models with "systemic risk".

Under the EU AI Act, GPAI models with systemic risk are those which meet any of the following criteria:[161]

a) The model has "high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks".
b) The European Commission has issued a decision identifying the model as having systemic risk based on criteria (a).
c) The model's "cumulative amount of computation used for its training measured in floating point operations is greater than 10^{25}".

Developers that have identified their model meets criterion (a) should notify the European Commission within two weeks. Decisions taken pursuant to a decision of the Commission in scenario (b) may be informed by the scientific panel established as part of the AI Act, which is made up of experts with "up-to-date scientific or technical expertise" and whose role includes the "development of tools and methodologies for evaluating capabilities"[162].

[157] AI Act Article 27. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[158] AI Act Article 3(63). Ibid.
[159] AI Act Article 53. Ibid.
[160] Under the AI Act, a "free and open licence" is one that "allows for the access, usage, modification, and distribution of the model"; and "parameters" includes its "weights, the information on the model architecture, and the information on model usage". AI Act Article 53. Ibid.
[161] AI Act Article 51. Ibid.
[162] AI Act Article 68. Ibid.

The AI Act requires developers of GPAI models with "systemic risk" to do the following:

- "perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identify and mitigate systemic risk";
- "assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, placing on the market, or use of [the model]";
- "keep track of, document and report without undue delay to [regulators] relevant information about serious incidents and possible corrective measures to address them";
- "ensure an adequate level of cybersecurity protection".

5.1.1.2 Analysis

While the AI Act marks a significant advancement in the regulatory oversight of AI, it falls short in addressing the complexities and risks associated with openness. The legislation's current framework does not fully tackle the challenges posed by the broad transparency necessary for effective external research, nor does it sufficiently promote the levels of openness required for comprehensive safety assessments. At its core, the AI Act emphasises developer accountability in ensuring the safety of their models; it places the onus on developers to determine the extent to which external researchers are involved in the safety and evaluation process. This approach could perpetuate the limitations in scope and transparency identified earlier, potentially compromising the overall safety of AI models.

Nevertheless, the AI Act does introduce elements of transparency that could benefit external researchers. The legislation mandates the maintenance of technical documentation, summaries of training data, and impact assessments related to fundamental rights. These requirements are designed to provide external researchers with a clearer understanding of model capabilities and limitations. Notably, the summaries of training data, which are expected to be publicly accessible, will offer a valuable resource. Although these summaries may vary in detail depending on implementation, the Act's provision for a standardised template aims to improve the consistency and utility of this information[163]. While this does not equate to full transparency, it represents a meaningful step forward in supporting both contracted and independent researchers in their efforts to assess model safety.

However, the AI Act does not mandate that fundamental rights impact assessments or technical documentation be made available to external researchers. Instead, developers are only required to share this information with regulators and business users (in the case of technical documentation). Accordingly, developers will be able to continue to withhold important information about the model, the risks identified by developers and the mitigation measures they have put in place. Furthermore, the scope of the fundamental rights impact assessments is limited, applying mainly to AI applications in public services, insurance and credit scoring.

[163] Tarkowski, A. (2024). "AI Act fails to set meaningful dataset transparency standards for open source AI." *Open Future.* https://openfuture.eu/blog/ai-act-fails-to-set-meaningful-dataset-transparency-standards-for-open-source-ai/; Warso, Z. & Keller, P. (2024). "Towards Robust Training Data Transparency." *OpenFuture.* https://openfuture.eu/publication/towards-robust-training-data-transparency/

The technical documentation requirements for GPAI models are similarly constrained by an open-source exception. As Downing notes, the AI Act "doesn't actually bother to define exactly what it means for models to be under "free and open source licenses"[164]. This ambiguity means that an exemption to the majority of the requirements applicable to GPAI could be achieved by releasing the model under a "open source" license which is nonetheless restrictive. It means that both licenses relying on the Open Source Initiative's open source AI definition[165] and the license used by Meta to release Llama 2 and 3[166], which puts restrictions on commercial uses for some users[167], could be used to this end. As Liesenfeld and Dingemanse explain, "what is to stop model providers from releasing the most inscrutable portion of their system (say, model weights) under an OSI-approved licence and collecting open source benefits? This stands to be a major avenue for open-washing."[168]

Though the AI Act's open source exemption does not apply to GPAI models classified as having "systemic risk", it still fails to ensure that evaluations of systemic risks are accessible to external researchers. This gap could hinder the Act's ability to prevent the release of models with significant risks, particularly if developers limit their risk assessments to a narrow set of concerns. For instance, a model might be released with limited external scrutiny, only for new risks to surface later that could have been identified through broader stakeholder involvement.

There are nonetheless opportunities to remedy this problem through "secondary legislation". Indeed, the AI Act provides that GPAI model developers (including those with "systemic risk") may rely on a Code of Practice, facilitated by regulators, which could address how and when external researchers should be involved in risk evaluation and mitigation. These Codes are expected within nine months of the Act's entry into force and could provide a framework for greater external involvement.

Another challenge with the AI Act is its approach to classifying GPAI models with systemic risk. The Act has a broad definition of "high impact capabilities" which is adaptable and allows the European Commission to make classifications informed by a scientific panel. However, the Commission lacks investigatory powers under the Act, and the scientific panel's ability to gain information from developers is similarly limited. This means that the onus is primarily on developers to designate their models as having systemic risk. If developers' commercial incentives prevail over the need for due diligence, they may be less likely to designate their models as such. Additionally, although the European Commission may unilaterally designate GPAI models as having systemic risk, there are no provisions in the AI Act which would enable it to make this decision before the model is actually released, including with a high degree of openness. Nor are there provisions that would see external stakeholders, whether researchers or regulators, being able to evaluate the

[164] Downing, K. (2024). "Choose Your Own Adventure: The EU AI Act and Openish AI." *Law Offices of Kate Downing.* https://katedowninglaw.com/2024/02/06/choose-your-own-adventure-the-eu-ai-act-and-openish-ai-2/

[165] The Open Source AI Definition – 1.0. *Open Source Initiative.* https://opensource.org/ai/open-source-ai-definition

[166] Llama 3. *Meta.* https://llama.meta.com/llama3/license/

[167] Maffulli, S. (2023). "Meta's LLaMa 2 license is not Open Source." *Open Source Initiative.* https://opensource.org/blog/metas-llama-2-license-is-not-open-source

[168] 168 Liesenfeld, A. & Dingemanse (2024). "Rethinking open source generative AI: open-washing and the EU AI Act." *FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency.* https://dl.acm.org/doi/10.1145/3630106.3659005: p3.

presence of systemic risk prior to release. As a result, a GPAI model may be released in a downloadable format, only for independent researchers and/or the Commission's scientific panel to identify the model as having systemic risk.

Another issue with the AI Act's classification of systemic risk is its reliance on a deterministic criterion linked to the amount of compute used in the training of the model. According to Bommasani, compute-based criteria for tiering models are "short-sighted", as the "relationship between compute and impact is quite tenuous and not evidentiated"[169]. Research institute Cohere for AI criticised the use of compute-based thresholds along similar lines[170]. Furthermore, the amount of compute necessary to train models in general is likely to change significantly over time. If it increases, a compute-based criteria runs the risk of applying to too many models, including those which do not carry significant risk. Conversely, if it decreases, the criteria may then not apply to any models at all. Accordingly, the AI Act includes a provision allowing the European Commission to adapt this threshold, which may result in more relevant criteria being used in the future.

5.1.2 The Digital Services Act

5.1.2.1 Description

The Digital Services Act (DSA) regulates the activities of intermediary services and online platforms, including 23 Very Large Online Platforms (VLOPs) and two Very Large Online Search Engines (VLOSEs)[171]. VLOPs and VLOSEs are services with at least 45 million active monthly users in the EU. The DSA regulates these services in areas that are directly relevant for Generative AI, such as content moderation, risk assessment and mitigation, and data access for researchers.

Under the DSA, vetted researchers have a right to request access to all relevant data of VLOPs and VLOSEs that could contribute to identifying and mitigating systemic risks. Under Article 40, platforms must provide vetted researchers with "access to data […] for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union […] and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures".

The framework for becoming a vetted researcher under the DSA is marked by strict criteria designed to ensure both the integrity of research and the protection of sensitive data. To qualify as a vetted researcher, several stringent requirements must be met: (i) being affiliated to a research organisation, whose definition potentially covers a wide range of non-university organisations, including civil society organisations; (ii) being independent from commercial interests; (iii) having an appropriate research proposal and disclosing the funding of the research; (iv) being able to fulfil data security and confidentiality requirements, as well as to protect personal data, and explain how they intend to do so. Vetted researchers must also commit to making their research results

[169] Bommasani, R. (2023). "Drawing Lines: Tiers for Foundation Models." *Stanford University.* https://crfm.stanford.edu/2023/11/18/tiers.html

[170] Cohere for AI (2024). "Exploring the Role of Compute-Based Thresholds for Governing the Risks of AI Models." https://cohere.com/research/papers/The-Limits-of-Thresholds.pdf

[171] Supervision of the designated very large online platforms and search engines under DSA. *European Commission.* https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses

publicly available free of charge, within a reasonable period after the completion of the research.

Applications for access to data must be sent to the national DSA regulator (the "digital services coordinators"). Platforms can ask for amendments to the request if they do not have the data or if they see security risks or risks to confidentiality, including trade secrets. Once granted, access must be provided to the researchers within a "reasonable period". VLOPs and VLOSEs are not required to remunerate the vetted researchers that gain access to their data, meaning research will be limited to the areas for which they able to secure funding.

5.1.2.2 Analysis

The DSA's access to data provision is a ground-breaking instrument in the sense that it is the first legal instrument that mandates VLOPs to provide access to relevant data to external researchers. However, as highlighted by Lemoine and Vermeulen[172], the DSA's scope means that this provision won't provide a direct backdoor through which researchers will be able to scrutinise Generative AI in general. It will only be relevant in two scenarios. First, Article 40 of the DSA applies to those Generative AI applications that can be qualified as "online search engines"[173] or "hosting" services[174] if and when those applications reach an average 45 million active monthly users in the EU. Second, Generative AI products, including LLMs, are also being embedded to enhance existing online services such as search engines (e.g. Bing Chat) and social media platforms (e.g. Snapchat's My AI). As such, Generative AI applications embedded within VLOPs or VLOSEs are covered by Article 40 of the DSA. All other Generative AI models and systems however fall outside of the scope of this provision.

5.2 The UK's approach

5.2.1 Description

The UK government has so far not introduced a bill on AI. In August 2023, the UK government published an AI White Paper setting out its principles-based approach for the governance of AI[175]. The focus is on empowering existing regulators to take responsibility for the establishment, promotion and oversight of responsible AI in their sectors rather than introducing new AI-specific legislation or new regulatory bodies. It tasks regulators with encouraging the adoption of ethical AI principles by UK industry, such as "safety, security and robustness" and "appropriate transparency and explainability".

[172] Lemoine, L. & Vermeulen, M. (2024). "Assessing the Extent to Which Generative Artificial Intelligence (AI) Falls Within the Scope of the EU's Digital Services Act: an Initial Analysis." *SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4702422

[173] Under the DSA, search engines are defined as "an intermediary service that allows users to input queries in order to perform searches of, in principle, all websites, or all websites in a particular language, on the basis of a query [...] and returns results in any format in which information related to the requested content can be found". Article 3(j) DSA. Regulation (EU) 2022/2065 on a Single Market For Digital Services (Digital Services Act). https://eur-lex.europa.eu/eli/reg/2022/2065/oj

[174] Under the DSA, a "hosting" service consists of "the storage of information provided by, and at the request of, a recipient of the service." Article 4(g) DSA. Ibid.

[175] UK DSIT and Office for Artificial Intelligence (2023) Policy paper: A pro-innovation approach to AI regulation. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

The AI White Paper also came with the establishment of the AI Safety Institute (UK AISI), a government research centre. Part of its remit includes undertaking external evaluations of foundation models. To this end, it secured voluntary commitments from OpenAI, Microsoft, Anthropic and Meta to conduct pre-release safety testing on their models in partnership with external researchers, such as AI evaluation organisations, with a focus on misuse and societal impacts, among other issues[176].

5.2.2 Limitations

The pre-release evaluation framework should provide committed developers with useful input for making their models safer, as will efforts to involve external researchers in those evaluations. However, its main weakness is linked to its voluntary nature. Indeed, despite their commitments, OpenAI, Anthropic and Meta have all released new models without going through a UK AISI evaluation first[177]. So far, Google's Gemini and Anthropic's Claude Sonnet 3.5[178] are the only models that have gone through an evaluation by the UK AISI, with the focus of the evaluations remaining unclear[179] .

According to reports, developers are hesitant to work with the UK AISI on pre-release evaluations out of a fear it may set a precedent in other jurisdictions[180]. Developers have also complained of a lack of transparency about how evaluations will be undertaken, their duration and its feedback process. The UK AISI's remit is highly limited in this respect, as the goal of evaluations is "not to designate any particular AI system as 'safe'", but simply to "enable better informed decision-making by governments and companies and act as an early warning system for some of the most concerning risks"[181]. As such, there is no mechanism that would prevent a company from releasing a model that the UK AISI has evaluated to be high risk.

Furthermore, although the UK AISI involves external organisations in its evaluations, it retains final say on the focus of evaluations. This decreases the benefits associated with bringing in external researchers, whose ability to use their diverse perspectives to guide research is therefore much more limited. Additionally, only organisations selected by the UK AISI may participate in its evaluations, which may decrease the diversity of external researchers involved.

[176] Milmo, D. & Stacey, K. (2023). "Tech firms to allow vetting of AI tools, as Musk warns all human jobs threatened." *The Guardian.* https://www.theguardian.com/technology/2023/nov/02/top-tech-firms-to-let-governments-vet-ai-tools-sunak-says-at-safety-summit ; UK DIST (2024). "AI Safety Institute: third progress report." *Gov.uk.* https://www.gov.uk/government/publications/uk-ai-safety-institute-third-progress-report/ai-safety-institute-third-progress-report

[177] Manacourt, V., Volpicelli, G. & Chatterjee, M. (2024). "Rishi Sunak promised to make AI safe. Big Tech's not playing ball." *Politico.* https://pro.politico.eu/news/178741

[178] The UK AISI is expected to be able to assess models by OpenAI and Anthropic in the future through an agreement with the US. Alder, M. (2024). "OpenAI, Anthropic enter AI agreements with US AI Safety Institute." *Fedscoop.* https://fedscoop.com/openai-anthropic-enter-ai-agreements-with-us-ai-safety-institute/

[179] Hern, A. (2024). "Claude 3.5 suggests AI's looming ubiquity could be a good thing." *The Guardian.* https://www.theguardian.com/technology/article/2024/jun/25/anthropic-claude-ai-chatbot

[180] Manacourt, V., Volpicelli, G. & Chatterjee, M. (2024). "Rishi Sunak promised to make AI safe. Big Tech's not playing ball." *Politico.* https://pro.politico.eu/news/178741

[181] UK Department for Science, Innovation and Technology (2024). "Introducing the AI Safety Institute." *Gov.uk.* https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

5.3 The US's approach

5.3.1 Description

The US's approach to Generative AI so far is mostly limited to the October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (AI Executive Order) [182]. It includes provisions on "dual-use foundation models", which broadly applies to Generative AI models[183].

The AI Executive Order furthermore has certain provisions relating to "dual-use foundation models with widely available weights", which loosely refers to Generative AI models whose weights are downloadable. To "address the risks and potential benefits of dual-use foundation models with widely available weights", the Secretary of Commerce undertook a consultation[184] on the "risks associated with actors fine-tuning […] or removing those models' safeguards", as well as "potential voluntary, regulatory and international mechanisms" to manage these risks while maximising the associated benefits of openness[185]. The consultation also considered the "benefits to AI innovation and research, including research into AI safety and risk management" of models with downloadable weights. The US government has not yet taken any action following the end of the consultation.

In connection with the AI Executive Order, the US National Institute of Standards and Technology (NIST) developed the Artificial Intelligence Risk Management Framework[186]. It contains recommendations for developers that could improve openness to external researchers, such as to publish the results of performance evaluations[187] and to collaborate with external researchers to maintain awareness of best practices and tools to measure and manage risks[188]. Following a consultation[189], the NIST has also published initial draft guidance on Managing Misuse Risk for Dual-Use Foundation Models[190]. Its recommendations may also improve openness to external researchers, such as establishing a safe harbour for independent researchers and a "program to incentivize researchers for finding vulnerabilities and disclosing them".

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

[183] Under the AI Executive Order, a "dual-use foundation model" is "an AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety, or any combination of those matters." Ibid.

[184] Dual Use Foundation Artificial Intelligence Models with Widely Available Model Weights. *NTIA.* https://www.ntia.gov/federal-register-notice/2024/dual-use-foundation-artificial-intelligence-models-widely-available-model-weights; NTIA AI Open Model Weights RFC. *NTIA.* https://www.regulations.gov/document/NTIA-2023-0009-0001

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

[186] US NIST (2024). "NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile." https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf

[187] MG-4.2-001. Ibid.

[188] MG-4.1-001. Ibid.

[189] US NIST (2023). "NIST Calls for Information to Support Safe, Secure and Trustworthy Development and Use of Artificial Intelligence." https://www.nist.gov/news-events/news/2023/12/nist-calls-information-support-safe-secure-and-trustworthy-development-and

[190] US AI Safety Institute (2024). "Managing Misuse Risk for Dual-Use Foundation Models (initial public draft)." https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf

The AI Executive Order requires Generative AI developers to report to the US Secretary of Commerce regarding (among other things) their training and development activities and the results of "AI red-team testing […] and a description of any associated measures the company has taken to meet safety objectives". The US government has additionally secured commitments from various Generative AI developers regarding the safe development and release of their models[191]. In this context, Amazon, Anthropic, Google, Inflection, Meta, Microsoft and OpenAI have committed to undertaking pre-release "internal and external security testing" and to "facilitat[ing] third-party discovery and reporting of vulnerabilities". The developers have committed to focusing in particular on risks such as "biosecurity and cybersecurity, as well as [Generative AI's] broader societal effects".

5.3.2 Limitations

Although still in its early stages, the US framework surrounding Generative AI has the benefit of actively tackling the issue of openness. Most importantly, the NIST Artificial Intelligence Risk Management Framework and the draft guidance on Managing Misuse Risk for Dual-Use Foundation Models both meaningfully consider external researchers' role in combatting the risks associated with openly releasing Generative AI models.

However, the US approach has two main limitations. The first is that it focuses too narrowly on "models with widely available weights". As discussed in section 3, weights are just one element of Generative AI models, and their accessibility alone is not sufficient for robust external research. Any framework that focuses too narrowly on this risks prioritising a degree of openness that is inadequate to the goals associated with external research. Secondly, the US approach is focused on guidance, standards and voluntary commitments, i.e. policy instruments that are ultimately non-binding. This may ultimately limit their effectiveness and uptake.

5.4 Assessment

Neither EU, UK or US frameworks adequately tackle the issue of openness in Generative AI models. Their frameworks either contain gaps in scope, are missing important provisions, or are voluntary in nature. Yet, the EU, UK and US initiatives together provide an outline for what a framework that adequately tackles the issue of openness could look like. Such a model would include the EU AI Act's focus on highly capable models and systemic risk, the EU DSA's framework for enabling external research and vetting researchers, the UK's focus on pre-release evaluations, and the US's focus on standards for red teaming and involving third parties in model development.

Accordingly, the next section will provide recommendations for what a policy framework that adequately balances the risks and benefits of openness should entail.

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

6 Policy Options to Balance the Risks and Benefits of Openness in AI regulation

While openness increases the ability of malicious actors to misuse models, remove model safeguards, and fine-tune models for specific harmful purposes, it also enhances the ability of external researchers to scrutinise these models, enabling them to identify risks and improve mitigations. In order to make the best use of this advantage in the openness challenge, it is important for policymakers to not address this challenge as a binary problem which requires separate rules for “closed” and “open” models. Instead, policy makers need to create tailored rules that reflect the various degrees of openness that model developers make available.

This section outlines a **policy framework for balancing the risks and benefits of openness**, summarised in **Annex 1**. The framework constitutes a series of complimentary policy options that could be implemented either through legislation or as part of non-legislative initiatives and standards. These include:

1. Threshold criteria for high-risk Generative AI models.
2. Standards for responsible release.
3. Systematic researcher vetting.
4. A safe harbour for independent researchers.
5. Subsidies for external research.
6. Standards on levels of access.
7. Due diligence requirements for model hosting platforms.

Figure 2 illustrates the framework by demonstrating how it intervenes at key points in the Generative AI process of development, release and moderation/maintenance. Many of these recommendations would greatly benefit from an international approach that aligned standards across jurisdictions. This is especially true for criteria defining high-risk models, standards for responsible release, and guidelines on access levels. Without such alignment, inconsistencies could undermine efforts to ensure AI safety globally, highlighting the need for a coordinated international response.

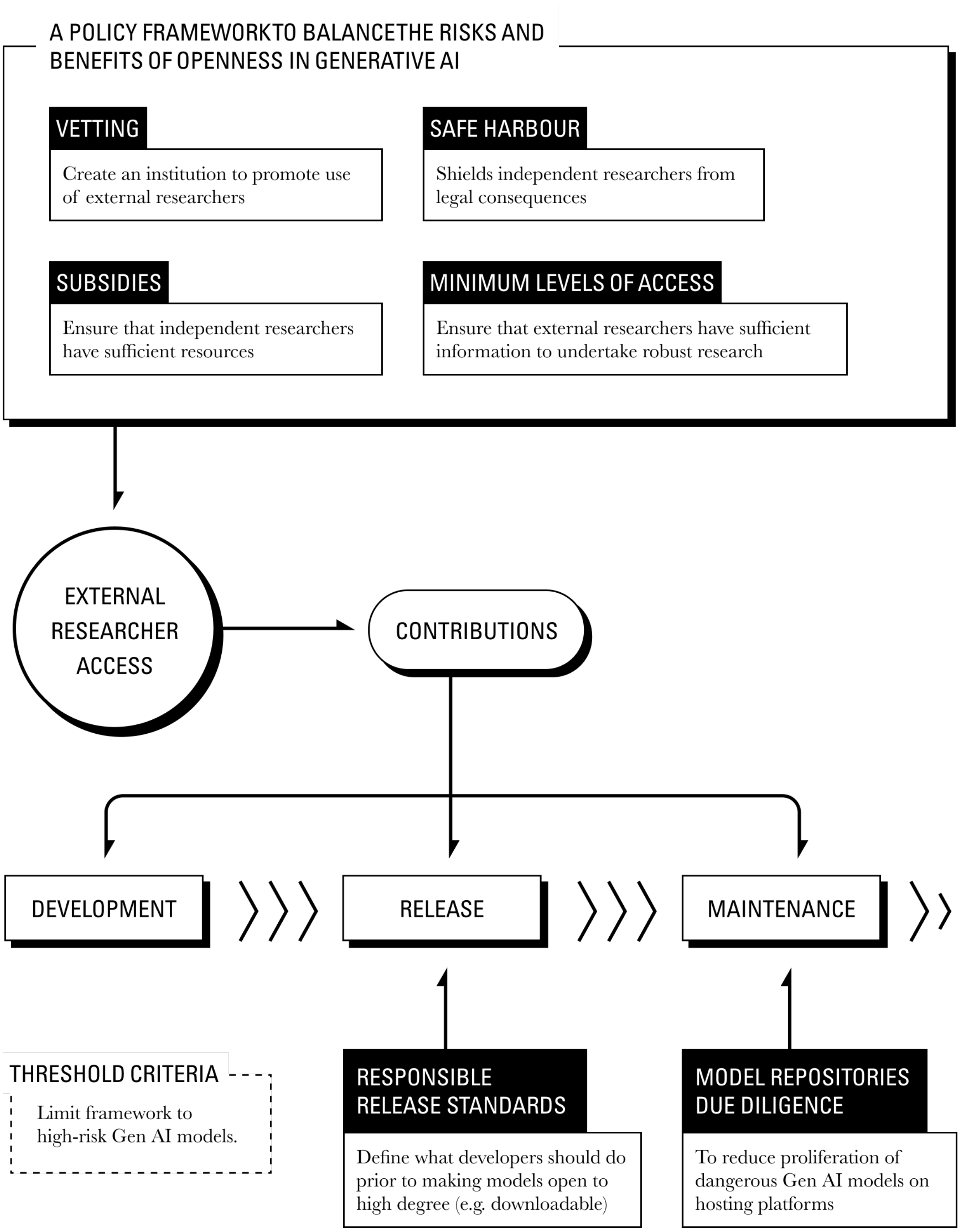

A POLICY FRAMEWORKTO BALANCETHE RISKS AND BENEFITS OF OPENNESS IN GENERATIVE AI
VETTING
Create an institution to promote use of external researchers
SAFE HARBOUR
Shields independent researchers from legal consequences
SUBSIDIES
Ensure that independent researchers have sufficient resources
MINIMUM LEVELS OF ACCESS
Ensure that external researchers have sufficient information to undertake robust research
EXTERNAL RESEARCHER ACCESS
CONTRIBUTIONS
DEVELOPMENT
RELEASE
MAINTENANCE
THRESHOLD CRITERIA
Limit framework to high-risk Gen AI models.
RESPONSIBLE RELEASE STANDARDS
Define what developers should do prior to making models open to high degree (e.g. downloadable)
MODEL REPOSITORIES DUE DILIGENCE
To reduce proliferation of dangerous Gen AI models on hosting platforms

FIGURE 2: ILLUSTRATION OF THE POLICY FRAMEWORK TO BALANCE THE BENEFITS AND RISKS OF OPENNESS IN GENERATIVE AI MODELS.

6.1 Threshold criteria for high-risk Generative AI models

The first step in any regulatory framework that seeks to address the challenges related to openness and Generative AI is identifying which models or categories of model require intervention. While numerous Generative AI models are developed every month, only some possess capabilities that could significantly propagate the risks identified in section 2. These high-risk models will require the most significant consideration by regulators, the most stringent legal requirements, and the most transparency. Additionally, clear criteria for high-risk models can help developers understand when to take additional care with the development of their models, such as by implementing extra safeguards, involving more external researchers or adopting a staged release strategy.

Therefore, developing criteria for high-risk Generative AI models is essential for focusing the efforts of both regulators and developers.

Developing risk criteria is, however, complex due to the wide variety of Generative AI models, their capabilities, use-cases and outputs. Deterministic criteria – such as the amount of compute used in training or the size of the developing organisation[192] – are seemingly objective yet imperfect tools. Nevertheless, they provide a useful starting point: deterministic thresholds could indicate the need for a more thorough investigation to determine if a model should be considered "high risk" or not. This investigation would then rely on harmonised methodologies and benchmarks to ascertain if a model merits special treatment.

Accordingly, a framework for determining threshold criteria for high-risk Generative AI models could consist of a combination of the following:

- **Deterministic thresholds:** Adaptable deterministic thresholds can signal the need for a proper risk evaluation prior to release. These criteria might include properties of the developing organisation (e.g., revenue) and/or the resources used to train the model (e.g., compute power).
- **Evaluations methodologies:** For models that meet the minimum deterministic threshold, a pre-release risk and mitigation evaluation should be conducted using standardised methodologies, indicators, and benchmarks. These standards should be adaptable over time, similar to the provisions in the EU AI Act.
- **Alerts:** Implementing a post-release alert system allows models identified as high risk by independent researchers or regulators to be labelled as such, ensuring ongoing monitoring and management of potential risks.

6.2 Standards for responsible release

A key objective of any policy framework for Generative AI is to prevent the release of high-risk models in a manner that could lead to widespread misuse, such as through non-gated downloadable access. This necessitates setting specific standards for responsible release for high-risk models.

Developing these standards is an iterative process that requires input from a diverse array of experts. Expertise in computer science, law, social science and safety is essential for assessing and updating the necessary safeguards within models over time. Once established, these standards could be enforced through regulation, holding developers liable if they release models without adhering to the responsible release guidelines.

A standard for responsible release could include:

- **External input:** Involving external researchers in evaluating and testing the model and its safeguards before release.
- **Best practice safeguards:** The implementation of certain key safeguards prior to release, which may include watermarking and traceability requirements, "poison pills" that disable the model if

[192] Bommasani, R. (2023). "Drawing Lines: Tiers for Foundation Models." *Stanford University.* https://crfm.stanford.edu/2023/11/18/tiers.html

someone tries to fine-tune for a particular purpose (e.g. creating CSAM), and child safety specific protections. It could also include data governance requirements, whereby a model's training data should not contain CSAM.

- **Governance**: Organisational measures that developers should have in place to ensure adequate model governance, such as quality management systems, recording and documentation procedures, adherence to best practices, and roles and responsibilities.
- **Staged release**: Releasing the model in stages to identify and mitigate emerging risks. Initially, the model could be made available through query API access to allow independent researchers to help identify risks before making the model fully downloadable.

6.3 Systematic researcher vetting

Creating a framework to systematically vet and accredit researchers could significantly boost both contracted and independent research in the field of Generative AI. By having researchers vetted, for example, by an independent institution, developers may be more likely to engage them during the development phase. Moreover, vetted researchers could gain access to certain model elements without necessarily needing a direct contract with the developers, such as technical documentation, training data, and the safety evaluations already undertaken by the developer. Additionally, vulnerability reports from vetted researchers could be prioritised in the developers' moderation activities.

It's however important that vetted research does not undermine other forms of research. While formal accreditation can foster trust and access, it should not exclude non-vetted researchers from contributing valuable insights. Maintaining a balance between vetted and non-vetted research will help ensure a diverse range of perspectives and expertise.

A framework to systematically vet researchers could involve the following:

- **Institutional framework:** Establish an institution – or empower an existing one – that is responsible for vetting researchers, including by overseeing a comprehensive application and review process.
- **Selection**: Implement stringent criteria for selecting researchers, covering aspects such as:
  - Affiliation with a research organisation.
  - Independence from commercial interests.
  - Adherence to data security and confidentiality requirements.
  - Disclosure of funding sources.
- **Accountability**: Ensure vetted researchers are held accountable, particularly regarding non-proliferation and the accuracy of vulnerability reports. This would include:
  - A complaints process for developers in cases where vetted researchers leak information, attempt to reverse engineer models, or otherwise breach their conditions of access.
  - Accountability for researchers issuing a high number of erroneous or inadequately-substantiated vulnerability reports.

Several researchers have proposed institutional frameworks for vetting, such as Liang and Bommasani's Foundational Models Review Board[193]. Harrington and Vermeulen have similarly proposed that developers create an independently-mediated structured researcher access programme, which would select researchers on their behalf[194]. Best practices in this area, and in responsible release generally, could also be specified by the scientific panel established under the AI Act[195]. As Harrington and Vermeulen suggest, this may need to be independently mediated and placed on a legislative basis to ensure developers do not maintain exclusive control to facilitate research that aligns with commercial incentives, drawing on the experience of the DSA[196].

6.4 Safe harbour for independent researchers

Independent researchers must be able to evaluate Generative AI models and systems without fear of legal consequences or liability for attempting to make models produce content that is illegal or forbidden by the developer's usage policy. To this end, it is essential that developers have a safe harbour for researchers that are not affiliated with the developer's organisation. By implementing a safe harbour, independent researchers would be encouraged to explore the boundaries and potential risks of Generative AI models without fear of legal repercussions. This would lead to a more robust understanding of the models' capabilities and vulnerabilities, ultimately contributing to the development of safer and more reliable AI technologies.

Accordingly, policymakers should encourage the adoption of safe harbours for independent researchers. Such safe harbours could include:

- **Liability exemptions**: Developers should not bring legal charges against researchers that disclose vulnerabilities and the presence of risks in good faith. This would involve delineating the specific circumstances and conditions under which independent researchers could attempt to breach an AI system's terms and conditions. This could make use of the vetted researcher process described in section 6.3, or delineate the specific technical and governance procedures that an organisation should have in place to prevent the illegal and harmful content from being used to non-research ends.
- **Moderation and appeals**: Developers should institute a moderation and appeals process that takes into account the activities of independent researchers[197]. When an independent researcher's account has been suspended for violating the model's usage policy, this decision should be sufficiently substantiated and include an appeals process. This appeals process could be fast-tracked for vetted researchers, researchers that

[193] Liang, P., Bommasani, R., Creel, K. & Reich, R. (2022). "The Time Is Now to Develop Community Norms for the Release of Foundation Models." *Stanford University.* https://crfm.stanford.edu/2022/05/17/community-norms.html

[194] Harrington, E. & Vermeulen, M. (2024). "External researcher access to closed foundation models." *Mozilla.* https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf

[195] AI Act Article 68. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[196] Harrington, E. & Vermeulen, M. (2024). "External researcher access to closed foundation models." *Mozilla.* https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf

[197] Longpre, S. et al. (2024). "A safe harbor for AI evaluation and red teaming." arXiv.org. https://arxiv.org/abs/2403.04893: p8.

have a researcher account, or simply researchers that are affiliated with a recognised research institution.

Longpre et al.'s seminal paper provides important guidance on the implementation of safe harbours for independent research[198]. It describes the procedure that could be implemented by developers to create a technical safe harbour, covering pre-registration processes for researchers, vulnerability reporting, the criteria for the application of a safe harbour, and the need for an impartial appeals process.

6.5 Subsidies for external research

As described in section 4.3, the quantity of external research undertaken is largely limited by the resources that developers spend on contracting external researchers and the funds available to independent researchers (e.g. public funds, philanthropy).

Accordingly, beyond increasing public funds available for AI research, policymakers could encourage developers to spend more on external researchers in the following ways:

- **Subsidised API access**: As described by Harrington and Vermeulen[199], developers could provide external researchers with subsidised access to their APIs. This could involve providing rate-limits that are sufficiently high to allow for automated evaluations. Policymakers could encourage developers to provide subsidised access schemes or establish independent intermediary bodies to provide subsidies on an application basis, or generally to vetted researchers (see section 6.3).
- **Tax rebates**: Developers could be provided tax rebates based on the amount of money that they spend on contracted researchers and subsidised API access more broadly.

6.6 Standards on levels of access

As described in section 3, the elements that developers make available to both contracted researchers and independent researchers vary and lack consistency. This can significantly affect the scope and quality of the research in question[200]. Policymakers could work to ensure greater levels of transparency for researchers by stipulating the minimum elements that should be released along with Generative AI models, and by facilitating greater degrees of access. This could involve the following:

- **Structured access**: Policymakers could encourage programs that provide external researchers with a differentiated and higher degree of access to Generative AI models. This could involve the use of modular API access, or "researcher API access", whereby researchers are provided access to the model through a special API that allows them to receive model outputs, examine model inputs, and even fine-tune the

[198] Ibid.
[199] Harrington, E. & Vermeulen, M. (2024). "External researcher access to closed foundation models." *Mozilla.* https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf
[200] Casper, S. et al. (2024). "Black-Box Access is Insufficient for Rigorous AI Audits." *arXiv.org.* https://arxiv.org/pdf/2401.14446

model (see section 3.1.2.3). Details of the level of access which modular API access should provide are described in great detail by Bucknall and Trager[201]. This higher degree of access could also be secured through onsite access, or gated downloadable access[202].

- **Minimum information**: Policymakers could specify the minimum categories of information that should be published along with Generative AI models. The AI Act, for example, requires general-purpose AI developers to publish detailed summaries of a model's training data[203]. Other information made available by default could include technical documentation and description of internal risk evaluations. Furthermore, standards could be established for how this information should be presented, and the level of detail[204].
- **Information security**: Policymakers should also encourage the uptake and development of information security procedures and tools that allow openness to external researchers while preventing model proliferation (e.g. through reverse engineering) and the leaking of trade secrets. Technical approaches include the use of federated learning[205] and the use of "fake" versions of the model[206]. Institutional approaches could involve allowing higher levels of access only to vetted researchers, as described in section 6.3.

6.7 Due diligence requirements for model hosting platforms

Despite ongoing policy and safety research aimed at making AI safer, a significant number of dangerous Generative AI models remain openly accessible. This is often due to inadequate evaluation prior to release or deliberate development and fine-tuning to propagate various risks. These models are frequently hosted on open-source platforms like GitHub, Hugging Face, and Civitai. As a result, there is some merit to exploring an international due diligence framework specifically for model hosting platforms. As proposed by Gorwa and Veale[207], such a framework could include:

- **Notice-and-action and takedown orders**: Asking hosting platforms to filter models at the point of upload would likely be too resource-intensive and onerous. However, a notice-and-takedown system that allows users to flag models that they have identified as dangerous would help hosting platforms to direct their attention where needed. Similarly, models that have been identified as disseminating illegal content by courts could be subject to takedown orders.
- **Pre-upload disclaimers**: Policymakers could explore the use of disclaimers for model developers that wish to make their models

[201] Bucknall, B. S. & Trager, R. F. (2023). "Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements." *University of Oxford.* https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements: P17.

[202] Open Minded (2023). "How to audit an AI model owned by someone else (part 1)." https://blog.openmined.org/ai-audit-part-1/

[203] AI Act Article 53. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[204] Warso, Z. & Keller, P. (2024). "Towards Robust Training Data Transparency." *OpenFuture.* https://openfuture.eu/publication/towards-robust-training-data-transparency/

[205] Bluemke, E., Collins, T., Garfinkel, B., Trask, A. (2023). "Exploring the Relevance of Data Privacy-Enhancing Technologies for AI Governance Use Cases." *ArXiv.org.* https://arxiv.org/abs/2303.08956

[206] Open Minded (2023). "How to audit an AI model owned by someone else (part 1)." https://blog.openmined.org/ai-audit-part-1/

[207] Gorwa, R. & Veale, M. (2024). "Moderating model marketplaces: platform governance puzzles for AI intermediaries." *Taylor & Francis.* https://www.tandfonline.com/doi/full/10.1080/17579961.2024.2388914

available for download. Uploaders could for example be required to certify that their model's training data does not contain CSAM and that their model's outputs contain a watermark (to aid traceability of content). While this might not deter all malicious actors, it could help hold them accountable once a model is identified as dangerous through takedown notices.

- **Know your model uploader**: Model hosting platforms could be required to collect certain identifiable information about the users that upload models to their platforms. This could for example include their name and address, and a copy of an identification document.

Acknowledgements

The authors would like to thank David E. Harris, Esme Harrington, Elizabeth Seger, Laureline Lemoine, Yisroel Mirsky, and Zoë Brammer for their valuable feedback, insightful comments and constructive suggestions during the drafting of this report.

This paper was commissioned by the Digital Infrastructure Insights Fund (D//F).

Annex 1: A Policy Framework to Balance the Risks and Benefits of Openness

Policy	*Goal*	*Method*
Threshold criteria for high-risk Generative AI models	Ensuring that any intervention is focused on only the most high-risk models.	• Deterministic thresholds (e.g. model size, compute used in training, size of developing organisation) used as starting point. • Pre-release risk evaluations for models that meet the deterministic threshold. • Post-release alert system for when researchers and regulators identify that a model is high risk
Responsible release standards	Prevent high-risk models from being released to a degree of openness which would allow their large-scale misuse (e.g. downloadable access).	Developers of high-risk models should: • Involve external researchers in pre-release evaluations for a minimum set of risk categories (e.g. AI-CSAM, disinformation, cyberattacks). • Implement a minimum set of safeguards (e.g. watermarking, clean training data) into high-risk models prior to release. • Implement a set a governance measures (e.g. quality management systems, documentation procedures). • Stage the release of their models (e.g. release through API prior to making downloadable).
Systematic researcher vetting	Creating an institution to promote developers' use of external researchers.	• Establishing an institution to vet researchers with an application and review process. • Stringent criteria for selecting researchers (e.g. affiliation with research institution, conflict of interest declarations, etc.). • Accountability mechanisms in cases where researchers leak confidential information or produce erroneous vulnerability reports.
Safe harbour for independent researchers	Ensuring that developers shield independent researchers from consequences for attempts to breach model usage policies and/or produce illegal content.	• Liability exemptions for external researchers focused on illegal content. • Developers should implement a clear and transparent moderation and appeals process for independent researchers.
Subsidies for independent research	Increase funding for independent research.	• Developers should provide subsidised API access to independent external researchers (e.g. to vetted researchers). • Tax rebates for the amounts spent on contracted researchers and subsidised API access.

Standards on levels of access	Ensuring that developers provide sufficient information to external researchers for them to undertake robust research.	• Encourage developers to implement programs whereby external researchers can be provided higher levels of model access ("structured access"). • Standards on the minimum categories of information that should be published along with Generative AI models (e.g. summary of training data). • Information security procedures and tools that allow openness to external researchers while preventing model proliferation and the leaking of trade secrets.
Due diligence requirements for model hosting platforms	Reduce the proliferation of dangerous Generative AI models on model hosting platforms.	• Notice-and-action frameworks on model hosting platforms, and a takedown order for models identified as dangerous by law enforcement or courts. • Pre-upload disclaimers for developers (e.g. certifying that model training data does not contain CSAM). • Model hosting platforms should be required to collect certain identifiable information about users that upload models.

Annex 2: Glossary

API: Application Programming Interfaces (APIs) are software interfaces that allow users to interact with Generative AI systems (e.g. ChatGPT).

Chatbot: Chatbots, or “conversational AI” allow the users to “speak” with Generative AI models using every-day language. Chatbots interpret instructions, questions and provide responses.

Contracted research: Research undertaken in exchange for remuneration on behalf of the developing organisation, usually for specific purposes such as security or bias research, and conducted through “structured access” methods tailored specifically for researchers.

CSAM: Child sexual abuse material (CSAM) is content that depicts acts of child sexual abuse and/or which focuses on the genitalia of children.

Dark web: Networks on the internet that are only accessible through special software, allowing users and operators to remain anonymous or untraceable. The dark web is not indexed by search engines, and therefore not readily accessible to most users.

Deepfake: Deepfakes are images or videos that have been digitally altered so that they appear to be someone else, typically for malicious purposes. Deepfakes can be non-consensual intimate images (NCII) (i.e. sexually explicit deepfakes created without the consent of the subject).

Fine-tuning: The process of adapting a pre-trained model for specific tasks or use cases, using data or training processes relevant to that use case.

Generative AI: Generative AI is a field of AI that focuses on creating new content based on existing data. Inputs and outputs of Generative AI include text, images, video and voice content.

Generative AI model: The base software and code that must be integrated within a system (e.g. an app) before it can be used.

Generative AI system: An application and software into which a Generative AI model has been integrated. It includes an API through which users can interact with the Generative AI model.

Generative AI companies: Companies that develop Generative AI models, for example Meta, Google, OpenAI, Stable Diffusion, Microsoft and Snapchat.

Independent research: Research undertaken on the basis of funding unrelated to the developing organisation, typically through “unstructured access” similar to that provided to general users.

Jailbreak: Also known as “direct prompt injection attacks”, jailbreaks involve engineering prompts to Generative AI models to circumvent safeguards put in place by developers in order to use the model in unintended ways.

LLM: Large Language Models (LLMs) are Generative AI models capable of interpreting and generating, along with other natural language processing tasks.

Red teaming: A way of interactively testing AI models to protect against harmful behaviour. Red teaming involves attempting to get the model to produce harmful content, as a means of testing the relevant safeguards.

Prompt: A prompt is the input that users give Generative AI models. This can include commands ("summarise this for me") and questions ("what is…?") provided through text or voice, and can be accompanied with other forms of media ("change this image").

External researcher: External research is conducted by the following groups: academics, research institutes, government agencies, industry bodies (e.g. established by industry for this particular purpose), non-profit organisations, and international organisations.

Model safeguards: Model safeguards are tools and structures put in place by developers that help ensure the model behaves in the intended manner. They can include, for example, input filters (preventing the use of certain inputs), watermarks and data governance practices.

Training data: The data used to train a Generative AI model. This data is typically scraped from the internet and social media platforms. It may include text, images, video and sound, and can include content representing children and children's data. Some training data sets are also available on an open-source basis for everyone to use.

VLOPs/VLOSEs: Under the EU DSA, Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) are online platforms and search engines that have a number of average monthly active users equal to or higher than 45 million.