A Safe Harbor for AI Evaluation and Red Teaming

Shayne Longpre* 1 Sayash Kapoor** 2 Kevin Klyman** 3 Ashwin Ramaswami 4 Rishi Bommasani 3
Borhane Blili-Hamelin 5 Yangsibo Huang 2 Aviya Skowron 6 Yong Zheng Xin 7 Suhas Kotha 8 Yi Zeng 9
Weiyan Shi 10 Xianjun Yang 11 Reid Southen 12 Alexander Robey 12 Patrick Chao 12 Diyi Yang 3 Ruosi Jia 9
Daniel Kang 13 Sandy Pentland 1 Arvind Narayanan 2 Percy Liang 3 Peter Henderson 2

March 5, 2024

Abstract

Independent evaluation and red teaming are critical for identifying the risks posed by generative AI systems. However, the terms of service and enforcement strategies used by prominent AI companies to deter model misuse have disincentives on good faith safety evaluations. This causes some researchers to fear that conducting such research or releasing their findings will result in account suspensions or legal reprisal. Although some companies offer researcher access programs, they are an inadequate substitute for independent research access, as they have limited community representation, receive inadequate funding, and lack independence from corporate incentives. We propose that major AI developers commit to providing a legal and technical safe harbor, indemnifying public interest safety research and protecting it from the threat of account suspensions or legal reprisal. These proposals emerged from our collective experience conducting safety, privacy, and trustworthiness research on generative AI systems, where norms and incentives could be better aligned with public interests, without exacerbating model misuse. We believe these commitments are a necessary step towards more inclusive and unimpeded community efforts to tackle the risks of generative AI.

1. Introduction

Generative AI systems have been deployed rapidly in recent years, amassing hundreds of millions of users. These systems have already raised concerns for widespread misuse, bias (Deshpande et al., 2023), hate speech (Douglas Heaven, 2020), privacy concerns (Carlini et al., 2021; 2023), disinformation (Burtell & Woodside, 2023), copyright infringement (Henderson et al., 2023; Gil et al., 2023), fraud (Stupp, 2019), weapons acquisition (Boiko et al., 2023; Urbina et al., 2022), and the proliferation of non-consensual and abusive images (Lakatos, 2023; Thiel et al., 2023), among others (Kapoor et al., 2024). To ensure sufficient public scrutiny and accountability, such high-impact systems should be evaluated (Liang et al., 2023; Solaiman et al., 2023; Weidinger et al., 2023) by independent and external entities (Raji et al., 2022; Birhane et al., 2024). Despite this, leading generative AI companies provide limited transparency and access into their systems, with transparent audits showing only 25% of policy enforcement and evaluation criteria were satisfied on average (Bommasani et al., 2023a); and with no company providing reproducible evaluations to characterize the effectiveness of their risk mitigations.

Leading AI companies’ terms of service prohibit independent evaluation into most sensitive model flaws (see Table 3). While these terms act as a deterrent to malicious behavior, they also restrict good faith research—auditors fear that releasing findings or conducting research could lead to their accounts being suspended, ending their ability to do such research, or even lawsuits for violating the terms of service. Already, in the course of conducting good faith research, researchers’ accounts have been suspended without warning, justification, or an opportunity to appeal (Marcus & Southen, 2024). While some companies authorize selected research through researcher access programs, their community representation remains limited and lacks independence from corporate incentives such as favoritism towards researchers aligned with the company’s values. Together, these observations stoke concerns that generative AI companies could emulate the transparency and accountability challenges with social media platforms—limiting researcher transparency and access can mitigate dangerous headlines, public relations fallout, and lawsuits, but at the expense of public interests (Abdo et al., 2022; DiResta et al., 2022).

As a group of researchers whose expertise spans AI red teaming, safety, and evaluation, as well as privacy, security, and the law, we have experienced first hand the negative
A safe harbor is a measure to provide legal protection to hackers engaged in "good faith" research, abiding by a company's terms of service. Violations of the usage policy are violations of the terms of service and can be enforced by terminating accounts or taking legal action.

Chilling effects describe the inhibition or discouragement of important research, in this case due to a lack of legal and technical protections, as well as uncertain norms around AI evaluation and red teaming.

### Table 1. We define and contextualize the technical terminology

<table>
<thead>
<tr>
<th>Terminology</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usage Policy</td>
<td>A company’s usage policy dictates what uses of its AI systems are acceptable or unacceptable. Usage policies generally prohibit inputs that elicit a range of undesirable model outputs, beyond what is already illegal. For example, see Anthropic’s Acceptable Use Policy.</td>
</tr>
<tr>
<td>Terms of Service</td>
<td>A company’s terms of service imposes legal rules on users of their services. Violations of the usage policy are violations of the terms of service and can be enforced by terminating accounts or taking legal action.</td>
</tr>
<tr>
<td>Generative AI Evaluation &amp; Red Teaming</td>
<td>In security fields, a red team refers to a group authorized to emulate an adversary’s attack against an organization’s security systems. This term has been adopted by the AI community to instead describe penetration testing of a broader set of system flaws than traditional security (The Hacking Policy Council, 2023). In this context, we are referring to testing of released systems by third party ethical hackers, who may or may have explicit consent.</td>
</tr>
<tr>
<td>Safe Harbor</td>
<td>A safe harbor is a measure to provide legal protection to hackers engaged in “good faith” research, abiding by pre-agreed rules of engagement, or vulnerability disclosure policy (e.g. HackerOne (2023)).</td>
</tr>
<tr>
<td>Good Faith Research</td>
<td>“Good faith security research means accessing a computer solely for purposes of good-faith testing, investigation, and/or correction of a security flaw or vulnerability, where such activity is carried out in a manner designed to avoid any harm to individuals or the public, and where the information derived from the activity is used primarily to promote security or safety...” (Department of Justice, 2022). We generalize this definition to research beyond security, including soliciting any unwanted behavior in the AI system normally disallowed by the company’s usage policy, which we broadly refer to as “safety research” in this work.</td>
</tr>
<tr>
<td>Vulnerability Disclosure Policy</td>
<td>A vulnerability disclosure policy establishes rules of engagement for third party ethical hackers. This includes disclosure requirements for discovered vulnerabilities, but also other mandatory protocols (Bugcrowd, 2023).</td>
</tr>
<tr>
<td>Chilling Effects</td>
<td>Chilling effects describe the inhibition or discouragement of important research, in this case due to a lack of legal and technical protections, as well as uncertain norms around AI evaluation and red teaming.</td>
</tr>
</tbody>
</table>

Table 1. We define and contextualize the technical terminology used in this work, which is often used in other disciplines.

### 2. Background & Motivations

Widely used online platforms can have significant socioeconomic impact (Zuboff, 2023; Horwitz et al., 2021). In this section we highlight three reasons to motivate new protections for independent research into generative AI platforms:

1. Social media research has been burdened by a lack of transparency and access, with a rise in legal repercussions for journalism and academic research (Abdo et al., 2022; DeLong, 2021; Belanger, 2023).
2. There is growing concern that widespread risks of generative AI will impact a wider swathe of society. Fostering wider participation in AI evaluation will require commitments to remove disincentives, obstacles, and favoritism in researcher access.
3. Independent AI evaluation is increasingly vital to fair assessments of AI risks, and informed policy debates.

We expand on each of these below, using terminology we define in Table 1.

#### 2.1. Avoiding the Fate of Social Media Platforms

Prominent social media platforms block researcher access to the detriment of public interests. Civil societies and researchers argue that social media companies have systematically limited researcher access to their platforms, restricting journalism and creating a chilling effect on critical public interest research (DiResta et al., 2022; Mozilla,
AI systems can generate toxic content (Deshpande et al., 2023), who accesses their systems, which could exacerbate gatekeeping (Li et al., 2023b; Pa Pa et al., 2023; Renaud et al., 2023), malicious tool usage (Casper et al., 2024). Today, AI systems like ChatGPT have amassed over 100 million weekly users (Hu, 2023), exceeding the number of Facebook users in 2022. Companies building generative AI models have the opportunity to protect good faith research before harm from their systems becomes as widespread as that from social media.

Conducting research on generative AI comes with additional challenges compared to social media. Compared to past digital technologies, prominent models require accounts to be used (unlike search engines), and their outputs are not publicly visible (unlike posts on many social media platforms) (Narayanan & Kapoor, 2023b). These factors provide developers with comparatively greater control over who accesses their systems, which could exacerbate gatekeeping. The lack of transparency from top developers compounds this issue, with little information available about how and where generative AI systems are used, and to what end (Bommasani et al., 2023b). For external researchers, the models themselves are also black boxes, as developers often do not disclose model architectures, sizes, or training data. This limits independent research to evaluate the risks, capabilities, safety, and societal impact of generative AI (Casper et al., 2024).

2.2. The Importance of Independent AI Evaluation

Concerns over the risks and harms of generative AI are mounting. Today, AI systems like ChatGPT have amassed over 100 million weekly users (Hu, 2023), exceeding the growth rate of social media platforms. Generative AI systems have already exhibited “unsafe” behavior—generating highly undesirable and even illegal content—attracting regulatory attention as a result. More specifically, generative AI systems can generate toxic content (Deshpande et al., 2023), libel, hate speech (Douglas Heaven, 2020), and privacy leaks (Carlini et al., 2021; 2023; Li et al., 2023a; Huang et al., 2023b; Nasr et al., 2023). They have also been used to scale disinformation (Burtell & Woodside, 2023), fraud (Stupp, 2019; Commission, 2023), malicious tool usage (Li et al., 2023b; Pa Pa et al., 2023; Renaud et al., 2023), copyright infringement (Henderson et al., 2023; Gil et al., 2023; Jonathan, 2023; Shi et al., 2024; Longpre et al., 2023), non-consensual intimate imagery (Lakatos, 2023), and child sexual abuse material (Thiel et al., 2023), as well as provide instructions for self-harm (Park et al., 2023; Xiang, 2023), acquiring weapons (Boiko et al., 2023; Nelson & Rose, 2023), and building weapons of mass destruction (Urbina, 2022; Soice et al., 2023). At the extreme end, even CEOs of AI model developers have speculated generative AI will upend labor markets (Suleyman & Bhaskar, 2023) and even pose more severe risks (Barrabi, 2023; Hendrycks et al., 2023). These wide ranging concerns, from the developers themselves, motivate the need for protected independent access.

Independent AI evaluation and red teaming are crucial for uncovering vulnerabilities, before they proliferate. Independent researchers often evaluate or “red team” AI systems for a broad range of risks. “Red teaming”, a subset of evaluation, has been adopted by the AI community as a term of art to describe these evaluations aimed at uncovering pernicious system flaws. In this work, we refer specifically to red teaming of publicly released AI systems (rather than pre-release testing), by external researchers, rather than internal teams. Some companies do also provide internal or by-invitation pre-release red teaming, e.g. OpenAI. While all types of testing are critical, external evaluation of AI systems that are already deployed is widely regarded as essential for ensuring safety, security, and accountability (Kenway et al., 2022; Anderljung et al., 2023). Post-release, external red-teaming research has uncovered vulnerabilities related to low resource languages (Yong et al., 2023), conjugate prompting attacks (Kotha et al., 2023), adversarial prompts (Maus et al., 2023; Zou et al., 2023; Robey et al., 2023), generation exploitation attacks (Huang et al., 2023a), persuasion attacks (Xu et al., 2023; Zeng et al., 2024), a wide range of jailbreaks (Wei et al., 2023; Shen et al., 2023; Liu et al., 2023; Zou et al., 2023; Shah et al., 2023), text-to-image vulnerabilities (Parrish et al., 2023), automatic red teaming (Ge et al., 2023; Yu et al., 2023; Chao et al., 2023; Zhao et al., 2024), and undetectable methods for fine-tuning away safety mitigations within the platform APIs (Qi et al., 2023; Yang et al., 2023; Zhan et al., 2023). See Appendix E for additional examples. These works illustrate how such research benefits AI companies: the research community assists in-house research teams by uncovering vulnerabilities, sharing findings and data, before systems cause major harm.

Independent AI evaluation provides impartial perspectives, that are necessary for informed regulation. As the above examples have shown, independent research has uncovered unexpected flaws, aiding company efforts, and expanding the collective knowledge around both vulnerabilities and defenses. These findings have informed the policy and regulatory discussions, including around the types of
We first discuss the mixed incentives and uncertainty faced by red teaming researchers, followed by analysis of the existing researcher protections, access programs, and their limitations.

### AI Companies’ Terms of Service discourage community-led evaluations

Many of the findings from the model vulnerability research mentioned in Section 2.2, such as jailbreaks, bypassing safety guardrails, or text-to-image exploits, are legally prohibited by the terms of service for popular systems, including those of OpenAI, Google, An-

<table>
<thead>
<tr>
<th>THEME</th>
<th>OBSERVATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chilling Effect on Research</td>
<td>Safety research can result in companies suspending researcher access, citing terms of service violations. This can have broad chilling effects as access is critical for the other work conducted by these researchers. Many researchers only begin their work after observing first-movers, and scope their practices to emulate those precedents. As a result, vital safety research may be delayed or circumscribed due to uncertainty and caution over account moderation outcomes.</td>
</tr>
<tr>
<td>Chilling Effect on Vulnerability Disclosure</td>
<td>It is unclear whether and how researchers should publicly release their findings, methods or the exploits themselves. In the absence of explicit guidance, they may be too broad or too limited in how they share their results, to the detriment of the community—for instance, by only sharing findings with a small group of other researchers, such as close personal contacts. When researchers are overly cautious in sharing their work it frequently results in siloed research that is less reproducible, or delayed disclosure, especially around the most sensitive findings, which could be to the detriment of public awareness.</td>
</tr>
<tr>
<td>Incentives to Tackle the Wrong Problems</td>
<td>There is an incentive to prioritize less important risks as the focus of safety work both for the uncertainty of repercussions from the companies or community. For instance, researchers might choose to investigate more benign prompt attacks rather than more offensive or dangerous attacks, such as focusing on text rather than more evocative visual outputs, or tool usage.</td>
</tr>
<tr>
<td>Favoritism and Imbalanced Representation</td>
<td>Admission into researcher access programs and favorable responses to safety work can be dependent on connections to the companies. For instance, there is a strong impression that access to OpenAI employees improves access to their programs. External researchers who are not already well connected may not hear back at all from their applications or receive any justification for rejection, as no obligation currently exists on the part of AI companies. Part of this may be due to companies being backlogged with applications from researchers, having dedicated few resources to this task. Costanza-Chock et al. (2022) point out the problems introduced by imbalanced auditor representation.</td>
</tr>
<tr>
<td>Unclear &amp; Undefined Norms</td>
<td>Impressions of basic norms and expectations vary widely, including with respect to appropriate threat models, whether and when to notify companies in advance of publication, what forms of red teaming are acceptable, and whether to release findings, methods, or prompts at all, or how to do so responsibly. Additionally, the type of API access, moderation policies, disclosure processes, and even the likelihood of response to a disclosure vary dramatically by company, leaving researchers without well-defined protocols that would enable them to confidently conduct important safety and security work.</td>
</tr>
<tr>
<td>A Choice Between Open and Closed Access</td>
<td>Researchers prefer to red team deployed systems that have millions of users and therefore pose immediate risks. However, effective and rigorous research requires deep access to the model (Casper et al., 2024), which proprietary systems rarely provide. As Friedler et al. (2023) have noted, “for red-teaming conducted by external groups to be effective, those groups must have full and transparent access to the system in question.” In particular, researchers often require finer-grained access to internal model representations (e.g., “logits”), access to both the base and aligned model, and continual access to a static model, without its API changing or becoming deprecated. Additionally, the underlying source of moderation in closed systems is difficult to diagnose: did the moderation endpoint catch an inappropriate user input, did the model itself abstain from answering, or did the user interface curtail an inappropriate response?</td>
</tr>
</tbody>
</table>

Table 2. Themes and observations attributed to informal discussions among authors and colleagues working on AI evaluation and red teeming. We describe the main challenges to conducting rigorous evaluations of widely used generative AI systems.
AI companies have begun using their terms of service to deter analysis, particularly into copyright claims. Midjourney updated its Terms of Service to include penalties such as account suspension or legal action for conducting such research.1 Midjourney’s Terms of Service states: “If You knowingly infringe someone else’s intellectual property, and that costs us money, we’re going to come find You and collect that money from You. We might also do other stuff, like try to get a court to make You pay our legal fees. Don’t do it” (Midjourney, 2023).2 Llama 2’s license will also terminate access if model outputs are used as part of intellectual property litigation.3

Our analysis of company policies in Table 3 shows not all companies disclose their enforcement process (the mechanisms for identifying and enforcing violations of the usage policy). Google and Inflection are the only companies to provide the user any form of justification on how the usage policy is enforced. And, only for OpenAI, Inflection, and Midjourney did we find evidence of an enforcement appeals process. Without additional information on how companies enforce their policies, researchers have no insight into enforcement appeals criteria, or whether companies reinstate public interest research post-hoc.

**Existing safe harbors protect security research but not other good faith research.** AI developers have engaged to differing degrees with external red teams and evaluators. OpenAI, Google, Anthropic, and Meta, for example, have bug bounties, and even safe harbors. However, companies like Meta and Anthropic currently “reserve final and sole discretion for whether you are acting in good faith and in accordance with this Policy”. They may revoke access rights to models, even open models like Llama 2 (Touvron et al., 2023), or hold the researchers legally accountable, at their discretion. This leaves clear ways to stifle and deter good faith research. Additionally, these safe harbors are tightly-scoped to traditional security issues like unauthorized account access.4 Developers disallow other model flaws named in their usage policies, including, “adversarial testing” (Anthropic, 2023), “jailbreaks”, bypassing safety guardrails, or generating hate speech, misinformation, or abusive imagery.

Among other safety research commitments, some companies publish reports on internal evaluation efforts, while others selectively invite third parties to participate in pre-release red teaming, or have researcher access programs for independent AI evaluation and red teaming.

**Independent AI evaluation is largely inconsistent, opaque, and challenging across companies.** From our experience and discussions, the bulk of this research is concentrated on Meta models like Llama-2 (Touvron et al., 2023), or OpenAI models like ChatGPT (OpenAI, 2023a). Llama models are popular as they have downloadable weights, allowing a researcher to red team locally without having their account terminated for usage policy violations. OpenAI models are popular as they are accessible via API, are highly performant, and have widespread public use. While many researchers are tentative about red teaming OpenAI, usage policy enforcement is often lax. However, account suspensions in the course of public interest research have taken place, to our knowledge, for each of OpenAI, Anthropic, Inflection, and Midjourney, with Midjourney being the most prolific. We withhold details on most of these to respect the anonymity of researchers. As one example, independent evaluation by an artist found Midjourney has a “visual plagiarism problem” (Marcus & Southen, 2024). This resulted in their account being repeatedly suspended without warnings or justification. The cost of suspensions without refunds quickly tallies to hundreds of dollars, and creating new accounts is also not trivial, with blanket bans on credit cards and email addresses.

AI companies have begun using their terms of service to deter analysis, particularly into copyright claims. Midjourney updated its Terms of Service to include penalties such as account suspension or legal action for conducting such research.1 Midjourney’s Terms of Service states: “If You knowingly infringe someone else’s intellectual property, and that costs us money, we’re going to come find You and collect that money from You. We might also do other stuff, like try to get a court to make You pay our legal fees. Don’t do it” (Midjourney, 2023).2 Llama 2’s license will also terminate access if model outputs are used as part of intellectual property litigation.3

Our analysis of company policies in Table 3 shows not all companies disclose their enforcement process (the mechanisms for identifying and enforcing violations of the usage policy). Google and Inflection are the only companies to provide the user any form of justification on how the usage policy is enforced. And, only for OpenAI, Inflection, and Midjourney did we find evidence of an enforcement appeals process. Without additional information on how companies enforce their policies, researchers have no insight into enforcement appeals criteria, or whether companies reinstate public interest research post-hoc.

**Existing safe harbors protect security research but not other good faith research.** AI developers have engaged to differing degrees with external red teams and evaluators. OpenAI, Google, Anthropic, and Meta, for example, have bug bounties, and even safe harbors. However, companies like Meta and Anthropic currently “reserve final and sole discretion for whether you are acting in good faith and in accordance with this Policy”. They may revoke access rights to models, even open models like Llama 2 (Touvron et al., 2023), or hold the researchers legally accountable, at their discretion. This leaves clear ways to stifle and deter good faith research. Additionally, these safe harbors are tightly-scoped to traditional security issues like unauthorized account access.4 Developers disallow other model flaws named in their usage policies, including, “adversarial testing” (Anthropic, 2023), “jailbreaks”, bypassing safety guardrails, or generating hate speech, misinformation, or abusive imagery.

Among other safety research commitments, some companies publish reports on internal evaluation efforts, while others selectively invite third parties to participate in pre-release red teaming, or have researcher access programs for independent AI evaluation and red teaming.
Table 3. A summary of the policies, access, and enforcement for major AI systems, suggesting a challenging environment for independent AI research. We catalog if each system has a public API, deeper access than final outputs (e.g. top-5 logits for OpenAI), researcher access programs, security research bug bounties, any legal safe harbors, and whether they disclose their account enforcement process, disclose justification on enforcement actions, and have an enforcement appeals process. • indicates the company satisfies this criteria; ○ indicates it does not, and ◆ indicates partial satisfaction. † Indicates security-only research safe harbors, “solely at [their] discretion”. ‡ Indicates a safe harbor for security and “academic research related to model safety”. The latter was added by OpenAI in response to reading an early draft of this proposal, though some ambiguity remains as to the scope of protected activities. Full details are provided in Table A1.

4. Safe Harbors

We believe that a pair of voluntary commitments could significantly improve participation, access, and incentives for public interest research into AI safety. The two commitments are: (i) a legal safe harbor, protecting good faith, public interest evaluation research provided it is conducted in accordance with well established security vulnerability disclosure practices, and (ii) a technical safe harbor, protecting this evaluation research from account termination; summarized in Figure 1. Both safe harbors should be scoped to include research activities that uncover any system flaws, including all undesirable generations currently prohibited by the usage policy. As we shall argue later, this would not inhibit existing enforcement against malicious misuse, as protections are entirely contingent on abiding by the law and strict vulnerability disclosure policies, determined ex post. Existing safe harbor resources (Etcovich & van der Merwe, 2018; Pfefferkorn, 2022; HackerOne, 2023), and vulnerability disclosure policies (Blog, 2010; Bugcrowd, 2023) provide grounding for these proposals. In particular, Elazari (2018b; 2019); Akgul et al. (2023); Kenway et al. (2022) discuss the implementations of algorithmic bug bounties, Walshe & Simpson (2023) note ambiguities on formal constraints, and Raji et al. (2022) explore governance for third-party AI audits, including legal protections for researchers. The legal safe harbor, similar to the proposal by Abdo et al. (2022) for social media platforms, would safeguard certain research from some amount of legal liability, mitigating the deterrent of strict terms of service and the threat that researchers’ actions could spark legal action by companies (e.g. under US laws such as the CFAA or DMCA...
Section 1201). The most important condition of a legal safe harbor is the determination of acting in good faith should not be “at the sole discretion” of the companies, as Meta and Anthropic have currently defined it. The technical safe harbor would limit the practical barriers erected by usage policy enforcement, with consistent and broader community access for important, public interest research. Together these steps would reduce the legal and practical obstacles to conducting independent evaluation and red teaming research.

4.1. A Legal Safe Harbor

A legal safe harbor could mitigate risks from civil litigation, providing assurances that AI platforms will not sue researchers if their actions were taken for research purposes. Take, for example, the U.S. legal regime, which governs many of the world’s leading AI developers. The Computer Fraud and Abuse Act (CFAA), which allows for civil lawsuits for accessing a computer without authorization or exceeding authorized access (CFAA, 1986), could be used by AI developers to sue researchers for accessing their models in a way that was unintended, though there are complexities to the legal analysis for adversarial attacks on AI models (Evitmov et al., 2019). Section 1201 of the Digital Millennium Copyright Act (DMCA) allows for civil lawsuits if researchers circumvent technological protection measures (TPMs), which effectively control access to works protected by copyright (DMCA, 1998a). These risks are not theoretical; security researchers have been targeted under the CFAA (Pfefferkorn, 2021), and DMCA § 1201 has deterred security researchers to the extent that they requested a DMCA exemption for this purpose (Colannino, 2021). Already, in the context of generative AI, OpenAI has attempted to dismiss the New York Times v OpenAI lawsuit (Grynbaum & Mac, 2023) on the allegation that New York Times research into the model constituted hacking (Brittain, 2024). Relatedly, a petition for an exemption to the DMCA has been filed requesting that researchers be allowed to investigate bias in generative AI systems (Weiss, 2023). Abdo et al. (2022) argue a safe harbor is oriented around conditions of access, rather than who gets access. The protections apply only to parties who abide by the rules of engagement, to the extent they can subsequently justify their actions in court. Typically, responsible vulnerability disclosure policies impose strict criteria for when the vulnerability should be disclosed, how long before it can be released to the public, privacy protection rules, and other criteria for the most dangerous exploits. Research that strays from those reasonable measures, or is already illegal, would not succeed in claiming those protections in an ex post investigation. As such, malicious use would remain legally deterred, and platforms would still be obligated to prevent misuse. Abdo et al. (2022) argue a safe harbor designed in this way, based on ex post researcher conduct, would not enable malicious use any more than in its absence. Nor would it alter platforms’ obligations to protect their users against third parties or from enforcing malpractice.

Companies’ legal safe harbors would protect researchers from civil liability, not criminal liability. Knowingly querying a model to generate certain types of content, whether for red teaming or not, can be illegal in certain jurisdictions—particularly in the case of image- or video-generation systems (Gupta, 2024). Moreover, certain violations of DMCA § 1201, particularly those that are committed “willfully and for purposes of commercial advantage or private financial gain,” can lead to criminal liability (DMCA, 1998b), as can many violations of the CFAA. We would recommend governments provide clear guidelines and, where appropriate, safe harbors for safe and responsible red teaming of illegal content generated by models. Such safe harbors against criminal conduct may need to be codified into statute in order to be guaranteed. However, they could be implemented by statements of policy, for example, such as when the Department of Justice issued a new policy in 2022 stating that “good-faith security research should not be charged” (Department of Justice, 2022).

The US Executive Order on AI directs the National Institute of Standards and Technology (NIST) to establish guidelines for conducting red-teaming and assessing the safety of foundation models (Executive Office of the President, 2023). Standardizing a legal safe harbor for researchers would complement NIST’s comprehensive AI evaluation agenda and its AI Risk Management Framework (NIST, 2024; Tabassi, 2023). The US AI Safety Institute Consortium, a public-private research collaboration, could be used to promote the adoption of safe harbors among companies (NIST, 2023).

4.2. A Technical Safe Harbor

Legal safe harbors still do not prevent account suspensions or other enforcement action that would impede independent safety and trustworthiness evaluations. Without sufficient technical protections for public interest research, a mismatch can develop between malicious and non-malicious actors since the latter are discouraged from investigating vulnerabilities exploited by the former. We propose companies offer some path to eliminate these technical barriers for good faith research. This would include more equitable opportunities for researcher access, and guarantees that those opportunities will not be foreclosed for researchers who adhere to companies’ guidelines.

The challenge with implementing a technical safe harbor is distinguishing between legitimate research and malicious actors, without notable costs to developers. An exemption to usage moderation may need to be reviewed in advance, or at least when an unfair account suspension occurs. However, we believe this problem is tractable, and offer recommen-
<table>
<thead>
<tr>
<th>Company Commitment: Legal Safe Harbor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commitment</strong> – We will not threaten or bring any legal action against anyone conducting good faith research who complies with the rules of engagement set out in our vulnerability disclosure policy. As long as you comply with our policy:</td>
</tr>
<tr>
<td>❖ We will not make any claim under the DMCA, for circumventing technological measures to protect the services eligible under this policy.</td>
</tr>
<tr>
<td>❖ We consider your security research to be “authorized” under the Computer Fraud and Abuse Act (and/or similar state laws).</td>
</tr>
<tr>
<td>❖ We waive any restrictions in our applicable Terms of Use and Usage Policies that would prohibit your participation in this policy, but only for the limited purpose of your model research under this policy.</td>
</tr>
<tr>
<td>❖ We will take steps to make known that you conducted good faith research if someone else brings legal action against you.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company Commitment: Technical Safe Harbor</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Commitment</strong> – We will make all reasonable efforts to not penalize user accounts engaged in good faith research into our systems, as long as they comply with the rules of engagement set out in our vulnerability disclosure policy.</td>
</tr>
<tr>
<td>❖ We shall not limit research on the basis that it may be against the interests of our company.</td>
</tr>
<tr>
<td>❖ We shall offer a research access program that involves independent, transparent, and timely review into research proposals.</td>
</tr>
<tr>
<td>❖ We shall offer a transparent appeals and review process if an account is restricted for alleged misuse (e.g. account suspension).</td>
</tr>
<tr>
<td>❖ We shall reinstate researchers’ accounts in the event that good faith research initiatives are found to have been penalized.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Good Faith Researcher Commitments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Scope of Research</strong> – Investigation into behavior of the AI system, including those disallowed by the acceptable usage policy.</td>
</tr>
<tr>
<td><strong>Researcher Responsibilities</strong> – All responsibilities, such as those already encoded in a company’s Rules of Engagement for security research continue to apply. These responsibilities include, but are not limited to:</td>
</tr>
<tr>
<td>❖ <strong>In-scope:</strong> Test only in-scope systems and respect out-of-scope systems.</td>
</tr>
<tr>
<td>❖ <strong>Vulnerability disclosure:</strong> Promptly report discovered vulnerabilities. Keep vulnerability details confidential if releasing them violates the law, or until a pre-agreed period of time after the vulnerability is reported (usually 90 days).</td>
</tr>
<tr>
<td>❖ <strong>Harms to users and systems:</strong> Refrain from violating privacy, disrupting systems, destroying data, or harming user experience.</td>
</tr>
<tr>
<td>❖ <strong>Privacy requirements:</strong> Do not intentionally access, modify, or use data belonging to others, including confidential data. If a vulnerability exposes such data, stop testing, submit a report immediately, and delete all copies of the information.</td>
</tr>
</tbody>
</table>

Figure 1. A summary of the suggested mutual commitments and scope of a legal safe harbor, and technical safe harbor. These commitments extend existing safe harbors for security research as well as researcher access programs, and are written in the context of US laws. For a wider list of common researcher responsibilities consider [OpenAI’s Rules of Engagement](https://www.openai.com/researchers/).
we propose would strengthen broader research protections. While this approach scales well and adopts independent access privileges, it can have severe limitations if AI companies only select a very finite set of partners, or choose to exclude more critical organizations. As a start, we recommend allowing NAIRR to help formulate the partner network, to include a set of trusted international academic organizations, as well as nonprofits in NAIRR such as AI2, EleutherAI, and MLCommons. Already these changes would make significant strides in expanding access through independent review.

**Transparent access and appeals processes can improve community trust.** Some generative AI companies may be unwilling to share access authorization more widely. There is a clear alternative: commit to a transparent access appeals process that makes decision criteria and outcomes visible to the wider community. Ideally, this process would be reviewed independently, perhaps with the help of NAIRR partner organizations. Whenever public interest evaluation research is suspended, researchers should have the opportunity to appeal the decision under a technical safe harbor. Companies can adopt an access process with clearly codified selection criteria, guaranteeing they will respond to applicants within a certain period of time, with a justification for the outcome decision. While this would not address the need for additional resources, it would provide the AI community with significantly greater visibility into companies’ decisions to grant access, and allow the community to apply collective pressure against any attempt to restrict legitimate research. The common denominator between pre-review and post-review technical safe harbors, described above, is providing a fair process to enable good faith research without the fear of unjustified account suspensions. In Appendix C we sketch an implementation of a pre-registration and appeals process, based on the existing researcher access programs, that could facilitate implementation of a technical safe harbor.

There are many dimensions of improving researcher access, including earlier access, deeper access, and subsidized access. The technical safe harbor described is a precondition for more independent and broader participation across all these axes, should companies offer earlier, deeper, or subsidized access. While efforts by AI companies to broaden safety research, such as accepting community applications for pre-release red-teaming and subsidizing such research with compute credits are useful first steps, the safe harbors we propose would strengthen broader research protections while being more independent of AI companies’ control.

5. **Related Proposals**

Our proposals for legal and technical safe harbors build on prior calls to expand independent access for AI evaluation, red teaming, and safety research. The Hacking Policy Council (2023) has proposed that governments “clarify and extend legal protections for independent AI red teaming,” similar to our voluntary legal safe harbor proposal. The Council stated, “the same industry norms on providing time to mitigate before public disclosure, and avoiding retaliation for good faith disclosures, should eventually apply to AI misalignment disclosures as they do for security vulnerability disclosures.” The Algorithmic Justice League has advocated for vulnerability disclosure for algorithmic harms, calling for independent algorithmic audits involving impacted communities (Costanza-Chock et al., 2022; Kenway et al., 2022). Moreover, AI Village hosts events where large groups of independent researchers red team generative AI models for a wide range of vulnerabilities (Sven Cattell, 2023). An array of researchers have recommended additional external scrutiny of the emerging risks and overall safety of frontier AI models to “improve assessment rigor and foster accountability to the public interest” (Anderljung et al., 2023). Bucknall & Trager (2023) have also proposed structured access for third party research via a dedicated research access API, with third-party independent review. Stanford’s Center for Research on Foundation Models has proposed an independent Foundation Models Review Board to moderate and review requests for deeper researcher access to foundation models (Liang et al., 2022).

Governments have also suggested the need for independent evaluation and red teaming. The US Office of Management and Budget’s Proposed Memorandum on Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence encourages federal agencies to consider as part of procurement contracts for generative AI systems “requiring adequate testing and safeguards, including external AI red teaming, against risks from generative AI such as discriminatory, misleading, inflammatory, unsafe, or deceptive outputs” (United States Office of Management and Budget, 2023). The EU AI Act states that providers of general-purpose AI models with systemic risks must share a “detailed description of the measures put in place for the purpose of conducting internal and/or external adversarial testing (e.g. red teaming), model adaptations, including alignment and fine-tuning” to the EU as part of their technical documentation (European Council, 2024; Hacker, 2023). In addition, Canada’s Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems includes a commitment that developers will “conduc[t] third-party audits prior to release” (Innovation, Science and Economic Development Canada, 2023).

6. **Conclusion**

The need for independent AI evaluation has garnered significant support from academics, journalists, and civil society.
Examining challenges to external evaluation of generative AI systems, we identify legal and technical safe harbors as minimum and fundamental protections. We believe they would significantly improve norms in the ecosystem and drive more inclusive community efforts to tackle the risks of generative AI.

Acknowledgements

We would like to thank Stephen Casper, Dylan Hadfield-Menell, Yacine Jernite, Amit Elazari, and Harley Geiger for their insightful feedback and guidance.

References


Hacker, P. Comments on the final trilogue version of the ai act, January 2023. URL https://media.lcdn.com/dms/document/media/D4E1PAQE9w01juCUvIw/feedsdocument-pdf-analyzed/0/1706022316786?e=1707350400&v=beta&t=POMy2m6nOFRLfHd4pO-Zj0JJWvehexHNLmWJlLYRA.


A Safe Harbor for AI Evaluation and Red Teaming


Pfaffkorn, R. Shooting the messenger: Remediation of disclosed vulnerabilities as cfca “loss”. Richmond
A Safe Harbor for AI Evaluation and Red Teaming


Appendix

A. Additional Considerations & Future Work

There are a number of future research directions that would help in making a safe harbor for AI evaluation and red teaming a reality. For instance, our proposal would benefit from further exploration of some of the challenging aspects in designing a technical safe harbor. In particular, to agree to such a commitment, AI companies will be concerned with protecting their own intellectual property and sensitive data. While restrictions on publicizing these valuable assets are often included in standard vulnerability disclosure policies, there is an implicit tension between expanding access to a greater number of independent researchers and ensuring compliance with disclosure policies. As AI models also expose new risks and harms, the definitions of “good faith” research may need to be flexible and evolve.

Our safe harbor proposals are formulated within the context of the US legal system. It is likely that different jurisdictions impose substantially different legal requirements related to research on the safety, security, and trustworthiness of AI. The use of geo-location in social media and search engines has allowed for digital platforms to tailor the behavior of their algorithmic systems based on each region. Generative AI companies may also adopt geo-location to customize their policies and enforcement of those policies by region. Future work should consider these changes and how a safe harbor proposal could work to achieve its aims in supporting fair, transparent, and inclusive good faith research internationally.

This line of research would also benefit from a more robust engagement with counterarguments to these proposals. While we believe the benefits of wider participation in independent AI safety and trustworthiness research will outweigh any risks to misuse, especially for well designed safe harbors, others may disagree. These trade-offs deserve more empirical analysis to understand the effects of such proposals.

B. Details on Access & Enforcement Policies

In Table 3 we summarize the policies, access, and enforcement for the major AI companies and their flagship systems. In Table A1 we link the evidence for each determination. And in this section we describe the criteria for each column in greater detail.

- **Usage Policy**: ● indicates that the company documents its acceptable usage policy, which they all do.
- **Deep Access**: ● indicates that the company provides some level of access to the AI system in question (OpenAI provides the top 5 logits and Meta provides open weights), ○ indicates there is no deeper access to the model (as is the case for all other companies).
- **Researcher Access**: ● indicates that the company maintains a researcher access program (OpenAI, or Meta with released model weights), ● indicates there is some access for researchers with some caveats (Anthropic has a limited early access program), ○ indicates there is no researcher access (as is the case for all other companies).
• **Safe Harbor:** ● would indicate that there is a legal safe harbor for model vulnerabilities beyond security research. ○ indicates there is form of commitment to research exemptions. OpenAI, Anthropic and Meta have a safe harbor for security research. ✗ indicates there is no safe harbor (all other companies). OpenAI’s new safe harbor (since updating in late January, in response to this proposal) is the closest to a full legal safe harbor, though there remains some ambiguity remains as to the scope of protected activities. For Cohere, while it does not have a safe harbor, their usage policy says “Note about adversarial attacks: Intentional stress testing of the API and adversarial attacks are allowable, but violative generations must be disclosed here, reported immediately, and must not be used for any purpose except for documenting the result of such attacks in a responsible manner.” Meta also provides a similar safe harbor for in-scope activities, which appear to be “integral privacy or security issues associated with Meta’s large language model, Llama 2, including being able to leak or extract training data through tactics like model inversion or extraction attacks.” However, like Anthropic, it’s safe harbor is determined at their sole discretion, and therefore provides limited benefit.

• **Enforcement process:** ● indicates that the company shares significant detail about how it enforces its usage policy such as the specific practices it uses for enforcement (OpenAI, Anthropic), ○ indicates there is little or no detail publicly available about the specific ways that the company enforces its usage policy (all other companies). Each company prescribes a prohibited set of uses, required by their terms of service, and all of these are enforced with moderation systems in the APIs and playgrounds, though only OpenAI and Anthropic openly disclose this. For instance, in QPT-4’s System Card OpenAI acknowledges using “a mix of reviewers and automated systems to identify and enforce against misuse”, and that policy-violating content will trigger warnings, suspensions and bans.

• **Enforcement justification:** ● would indicate that the company provides a specific reason for why a certain prompt or query was violative, ○ indicates that the company provides some detailed (if non-specific) justification when a user’s prompt or query is blocked otherwise deemed violative (Google, Inflection). ✗ indicates there is no significant justification provided (all other companies).

• **Enforcement Appeal:** ● indicates that the company provides an appeals process when it takes an enforcement action under its usage policy (OpenAI, Inflection, Midjourney). ○ indicates there is no appeals process (all other companies).

### C. Implementation of a Technical Safe Harbor

In Section 4.2 we discuss two approaches by which companies can establish a technical safe harbor—by scaling researcher participation and enlisting independent judgement of what constitutes good faith research, without taxing corporate resources. These approaches offer two lenses: pre-review of research applications or post-review of suspended researchers. In reality, some combination of the two may be most convenient and efficient. Here we sketch a proposal for an independently reviewed appeals process (post-review), but that requires research pre-registration to ease the challenge of reviewing whether research is good faith. A key choice is to determine the set of acceptable institutions for research pre-registration, which would ideally be negotiated ahead of time with NAIRR. We sketch what the components of this system might look like:

• **Good Faith Research Pre-Registration:** Good faith researchers can pre-register their work, establishing in advance their affiliations, intent, and research goals, so the company can easily cross-reference flagged accounts with these detailed forms. Similar to the existing OpenAI Researcher Access Program, or Twitter’s 2021 Researcher API (before it was decommissioned), the pre-registration form can include: Name, API key, institutional affiliations, evidence of affiliation (email and website), list of investigators, intended research focus, specific sensitive topics that violate the usage policy, timeline, etc.

• **Vulnerability Disclosure:** The researchers should tag vulnerability disclosures through the same platform, so these can be directly connected to the pre-registration form.

• **Criteria for Technical Safe Harbor:** If an account is flagged to a company, either because it violated its usage policy, or for some other reason, the company can directly cross-reference the account with pre-registered forms. If a pre-registration does not exist, the company can suspend the account. If a pre-registration form does exist, the company can review the account’s eligibility for an exemption from enforcement based on a number of factors: (i) is the account affiliated with a recognized academic or research institution, (ii) are the usage policy violations in line with the proposed research topics/timeline, and (iii) is there any evidence that the researcher has violated the vulnerability disclosure policy, such as publishing vulnerabilities without advance disclosure (in the required timeframe). We recommend that acceptable research institutions be negotiated in advance under the guidance of NAIRR. Ideally the group of acceptable research institutions would include major international universities as well as organizations with a track record for
trusted research, such as AI2, EleutherAI, and Masakhane. In the event that each of these criteria are met and the
ccompany still has concerns, it can suspend the account and then directly contact the organization or supervisor of the
work, as disclosed in the form, with the justification for suspension.
• Suspension Appeals Process: If the account is suspended, despite the researcher having pre-registered their research
plan, there may be an incongruity or ambiguity in their application. The account holder will have the option to appeal
this process, ideally with an impartial, independent reviewer. If necessary, the company could escalate the appeal to
the university or organization’s department leads, to ensure the organization stands by the researcher’s work. This
would likely rule out the vast majority of malicious actors, and distribute the responsibility between AI companies and
research institutions themselves. The appeals process should have standardized, well-documented criteria and a fair
timeline (e.g. 30 days).

D. Company Support for Wider Participation in AI Evaluations

There is ample evidence that prominent AI companies are verbally committed to independent and broader AI system
evaluations. OpenAI’s Sharing & Publication Policy states “we believe it is important for the broader world to be able to
evaluate our research and products, especially to understand and improve potential weaknesses and safety or bias problems
in our models” (OpenAI, 2023b). It remains unclear how this commitment relates to OpenAI’s terms of service and their
enforcement.3 Anthropic has stated in its Core Views on AI Safety that “in the near future, we also plan to make externally
legible commitments to only develop models beyond a certain capability threshold if safety standards can be met, and to
allow an independent, external organization to evaluate both our model’s capabilities and safety” (Anthropic, 2023a). As part
of its Secure AI Framework, Google has committed to “Expanding our bug hunters programs (including our Vulnerability
Rewards Program) to reward and incentivize research around AI safety and security” (Hansen & Venables, 2023). Meta has
highlighted the importance of external red teams in improving the safety of Llama 2, noting that “Our extensive testing
through both internal and external red teaming is continuing to help improve our AI work across Meta” (Meta, 2023). In the
same vein, Inflection states “Red-teaming is and will continue to be the engine at the heart of our evaluation framework.
Red-teams provide the best indication of how a model will perform in real-world situations … To do this, we commission
outside experts as well as relying on our safety team. Inflection is currently building teams of highly specialized red-teamers
that can bring their unique expertise to investigate models in a manner our ‘in-house’ teams would not have the context to do
effectively” (Inflection, 2023). The Frontier Model Forum, comprised of OpenAI, Google, Anthropic, and Microsoft, states
that one of its core objectives is “Advancing AI safety research … Research will help promote the responsible development
of frontier models, minimize risks, and enable independent, standardized evaluations of capabilities and safety.”

E. Additional Red Teaming Work

In addition to the works on AI audits, red teaming, and evaluations cited in Section 2, there are many other notable works,
worthy of further discussion. Shah et al. (2023) find GPT-4 will give instructions for making weapons and narcotics. Fang
et al. (2024) shows how GPT-4 can be used to automatically hack websites in the right circumstances. Sharma et al. (2023)
discuss the behavior of model sycophancy. Santurkar et al. (2023) show political and ideological biases systemic in AI
models. Ji et al. (2023); Zhang et al. (2023) demonstrate the challenges with model hallucination. Qu et al. (2023) illustrates
models’ capacities for harmful content generation. Lastly, Rando et al. (2022) red teams Stable Diffusion’s safety filters,
revealing flaws.

3We emailed “papers@openai.com” to ask for clarification on research exemptions for the OpenAI Usage Policy, but received no
response.