Position: 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] **Zheng-Xin Yong**[7] **Suhas Kotha**[8] **Yi Zeng**[9] **Weiyan Shi**[10] **Xianjun Yang**[11] **Reid Southen** **Alexander Robey**[12] **Patrick Chao**[12] **Diyi Yang**[3] **Ruoxi Jia**[9] **Daniel Kang**[13] **Sandy Pentland**[1] **Arvind Narayanan**[2] **Percy Liang**[3] **Peter Henderson**[2]

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 generative AI developers commit to providing a legal and technical safe harbor, protecting public interest safety research and removing 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]MIT [2]Princeton University [3]Stanford University [4]Georgetown University [5]AI Risk and Vulnerability Alliance [6]Eleuther AI [7]Brown University [8]Carnegie Mellon University [9]Virginia Tech [10]Northeastern University [11]UCSB [12]University of Pennsylvania [13]UIUC. [*] Corresponding author. [**] Equal contribution. Correspondence to: Shayne Longpre <slongpre@media.mit.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Generative AI systems have been deployed rapidly in recent years, amassing hundreds of millions of users. These systems have already raised concerns of widespread misuse, bias (Deshpande et al., 2023), hate speech (Douglas Heaven, 2020), privacy (Carlini et al., 2021; 2023), disinformation (Burtell & Woodside, 2023), self harm (Park et al., 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; Toner & McCauley, 2024). However, (Bommasani et al., 2023a) found that leading generative AI companies provide limited transparency and access into their systems. In particular, they found that only 25% of policy enforcement and evaluation criteria were satisfied on average, and no company provided reproducible evaluations to characterize the effectiveness of their risk mitigations.

Leading AI firms' 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 result in their accounts being suspended, ending their ability to do such research, or even lawsuits for violating such terms. In the process of conducting good faith research, researchers' accounts have been suspended without warning, justification, or an opportunity to appeal (Marcus & Southen, 2024). Some companies authorize selected research through researcher access programs, but their community representation remains limited and lacks independence from corporate incentives such as favoritism towards researchers aligned with the companies' values. Taken together, these observations exacerbate concerns that generative AI companies could suffer similar challenges to social media companies in

TERMINOLOGY	CONTEXT
Acceptable Use Policy	A company's acceptable use policy dictates what uses of its AI systems are acceptable or unacceptable (Klyman, 2024). Acceptable use 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.
Terms of Service	A company's terms of service imposes legal rules on users of their services. Violations of the acceptable use policy are violations of the terms of service and can be enforced by terminating accounts or taking legal action.
Generative AI Evaluation & Red Teaming	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.
Safe Harbor	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)).
Good Faith Research	"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 acceptable use policy, which we broadly refer to as "safety research" in this work.
Vulnerability Disclosure Policy	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).
Chilling Effects	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** used in this work, which is often used in other disciplines.

terms of transparency and accountability, where companies have limited researcher access to reduce negative headlines and lawsuits at the expense of the public interest (Cattell & Ghosh, 2024; 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 law, we have experienced firsthand the negative effects of legal uncertainty and technical barriers to conducting important research (Table 2). To improve the status quo, **in this paper, we argue that generative AI companies must commit to two protections for independent public interest research**. First, AI companies should provide a **legal safe harbor** by offering legal protections for good faith research, provided it is conducted in line with vulnerability disclosure policies (as defined in Table 1). Second, companies should provide a **technical safe harbor**, protecting safety researchers from having their accounts subject to moderation or suspension. These are fundamental access requirements for inclusive evaluation of generative AI systems. Building on prior work for algorithmic bug bounties (Elazari, 2018a; Kenway et al., 2022; Raji et al., 2022) and social media data access (Abdo et al., 2022), we recommend ways to implement these protections for independent generative AI evaluation without undermining the processes that prevent model misuse. Specifically, we propose that companies delegate account authorization to trusted universities or nonprofits, or provide transparent recourse for accounts suspended in the course of research. These voluntary commitments align with the stated goals of AI companies: to support wider participation in AI safety research, minimize corporate favoritism, and encourage community safety evaluations (see Appendix D). We hope generative AI companies will adopt these commitments to establish better community norms, improve trust in their services, and bolster AI safety in proprietary systems.[1]

2. Background & Motivations

Widely used online platforms can have significant societal 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 risks from generative AI will impact much of society. Fostering wider participation in AI evaluation will require firms to remove disincentives 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.

[1]While our focus in this paper is generative AI, we recognize that protections and support for researcher access and audits are also essential for other types of AI systems and digital platforms.

2.1. Avoiding the Fate of Social Media Platforms

Prominent social media platforms block researcher access to the detriment of public interests. Civil society 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, 2023; Boyd et al., 2021; Persily, 2021). Specifically, platforms wield their terms of service to gatekeep access to publicly posted data and limit negative exposure from independent research. Abdo et al. (2022) argue for "a safe harbor for platform research," which would include legal provisions that protect researchers and journalists. In the absence of such provisions, researchers have reported platform gatekeeping, account suspensions, cease-and-desist letters and general fears of liability in the course of public interest research, which has had a chilling effect (DeLong, 2021; Barclay, 2021; Belanger, 2023). Academics in the fields of computer and internet security have also been subject to legal threats and lawsuits (Greene, 2001; Brodkin, 2021; disclose.io, 2021), resulting in new guidelines from the U.S. Department of Justice stating that "good-faith security research should not be charged" (Department of Justice, 2022). Firms 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). Generative AI systems have already exhibited problematic behavior—generating undesirable and even illegal content—attracting the attention of regulators. 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 et al., 2022; Soice et al., 2023). CEOs of AI companies have speculated that generative AI will upend labor markets (Suleyman & Bhaskar, 2023) and even pose more extreme risks (Barrabi, 2023; Hendrycks et al., 2023). These wide ranging concerns, including from AI companies, 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. 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; Raji et al., 2022). 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 perspec-

THEME	OBSERVATIONS
Chilling Effect on Research	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.
Chilling Effect on Vulnerability Disclosure	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.
Incentives to Tackle the Wrong Problems	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.
Favoritism and Imbalanced Representation	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.
Unclear & Undefined Norms	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.
A Choice Between Open and Closed Access	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?

Table 2. **Themes and observations attributed to informal discussions among authors and colleagues working on AI evaluation and red teaming.** We describe the main challenges to conducting rigorous evaluations of widely used generative AI systems.

tives necessary for informed regulation. As the above examples have shown, independent research has uncovered unexpected flaws, aiding company efforts, and enhanced collective knowledge about both vulnerabilities and defenses. These findings have informed policy and regulatory discussions, including around the types of model vulnerabilities, and the relative safety of open and closed foundation models (Narayanan & Kapoor, 2023a; Lambert, 2023). However, it is not evident that we are seeing the full benefits from a thriving red teaming ecosystem (Section 3).

Without robust independent evaluation, companies' own developer safety teams may not be sufficiently large or diverse to fully represent the diversity of global users their products already serve, and the scale of risks they have acknowledged (Costanza-Chock et al., 2022). While companies do invite third-party evaluators, there are well known conflicts of interest without independence in the auditor selection process (Moore et al., 2006). As the Ada Lovelace Institute and another dozen civil societies remarked at the recent AI Safety Summit in the UK, "Companies cannot be allowed to assign and mark their own homework. Any research efforts designed to inform policy action around AI must be conducted with unambiguous independence from industry influence" (Ada Lovelace Institute, 2023).

3. Challenges to Independent AI Evaluation

We first discuss the disincentives and uncertainty faced by independent researchers, followed by analysis of existing researcher protections, access programs, and their limitations.

Firms' 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, Anthropic, Inflection, Meta, Midjourney, and others. While these terms are intended as a deterrent against malicious actors, they also

inadvertently restrict safety and trustworthiness research. While platforms enforce these restrictions to varying degrees, these terms disincentivize good faith research by granting developers the ability to terminate researchers' accounts (without appeal or justification) or even take legal action against them. The risk of losing account access may dissuade many researchers altogether, as their accounts are critical for a range of vulnerability and other AI research.

AI developers' documentation often purports to support independent research; however, it does not clearly state the conditions under which evaluation and red teaming would not violate acceptable use policies, leaving researchers uncertain as to whether or how they should conduct their research. In Table 2, we share common themes attributed to discussions between ourselves and colleagues, summarizing their experiences conducting evaluation and red teaming research on generative AI platforms. These themes reflect an imperfect sample: they are skewed in that they represent the opinions of researchers *who chose to conduct safety research*, excluding those who chose not to, lacked access to the companies they would have evaluated, or were deterred by potential legal liability.

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 acceptable use 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, anecdotally, their enforcement of the acceptable use policy is 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.

Firms have begun using their terms of service to deter analysis, in particular related to copyright. Midjourney updated its terms of service to include penalties such as account suspension or legal action for conducting such research.[2] Midjourney's terms of service state: "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, Section 10). 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 acceptable use policy). Google and Inflection are the only companies to provide the user any form of justification on how their acceptable use policies are 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 criteria for appeals, 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 teamers 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 allows firms to stifle and deter good faith research. Existing safe harbors are tightly-scoped to traditional security issues like unauthorized account access.[4] Developers exclude other model flaws named in their acceptable use 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 deeper access to released models. These are laudable initiatives, especially when they are accompanied by subsidized credits for researchers, as the cost of model access and use can be a major hindrance for independent researchers (OpenAI, 2024b). Nonetheless, these measures leave significant gaps in the ecosystem for independent evaluations. Reports on internal red teaming are often largely irreproducible and

[2] See https://twitter.com/Rahll/status/1739155446726791470

[3] See Section 5c of the Llama 2 license (which Llama 3 shares).

[4] OpenAI expanded its safe harbor to include "model vulnerability research" and "academic model safety research" in response to an early draft of our proposal, though some ambiguity remains as to the scope of protected activities (OpenAI, 2024a).

AI Company	AI System	Public API / Open	Deep Access	Researcher Access	Bug Bounty	Safe Harbor	Enforcement Process	Enforcement Justification	Enforcement Appeal
OpenAI	GPT-4	●	◑	●	●	◑[†]	●	○	◑
Google	Gemini 1.0	●	○	○	●	○	○	◑	○
Anthropic	Claude 2	○	○	◑	○	◑[‡]	●	○	○
Inflection	Inflection-1	○	○	○	○	○	○	◑	◑
Meta	Llama 2	●	●	●	●	◑[‡]	○	○	○
Midjourney	Midjourney v6	○	○	○	○	○	○	○	◑
Cohere	Command	●	○	●	○	◑	○	○	○

Table 3. **A summary of the policies, access, and enforcement for major AI systems.** 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.

generate limited trust due to mismatched corporate incentives (e.g. Anthropic (2023b)). Invitations to third-party researchers are limited and can be self-selecting. And researcher access programs, if available, often do not notify researchers of rejections and thus create an environment of uncertainty (Bommasani et al., 2023b). Researchers have argued that a patchwork of policies like these can create a veneer of open and responsible research, without lifting other obstacles for participatory research (Krawiec, 2003; Zalnieriute, 2021; Whittaker, 2021).

Companies should take steps to facilitate independent AI evaluation and reduce the fear of reprisals for safety research. The gaps in the policy architectures of leading AI companies, depicted in Table 3 force well-intentioned researchers to either wait for approval from unresponsive access programs, or risk violating company policy and potentially losing access to their accounts. The net result is a situation akin to companies gatekeeping access to their platforms and thereby restricting the scope of safety research, whether intentional or not. This research environment can limit the diversity and representation in evaluation, ultimately stymieing public awareness of risks to AI safety.

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 acceptable use 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 following 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) (Klyman et al., 2024). 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 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 system, which governs many of the world's leading AI developers. The

Company Commitment: Legal Safe Harbor

Commitment – 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:

- We will not make any claim under the DMCA, for circumventing technological measures to protect the services eligible under this policy.
- We consider your security research to be "authorized" under the Computer Fraud and Abuse Act (and/or similar state laws).
- We waive any restrictions in our applicable Terms of Use and Acceptable Use Policies that would prohibit your participation in this policy, but only for the limited purpose of your model research under this policy.
- We will take steps to make known that you conducted good faith research if someone else brings legal action against you.

Company Commitment: Technical Safe Harbor

Commitment – 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.

- We shall not limit research on the basis that it may be against the interests of our company.
- We shall offer a research access program that involves independent, transparent, and timely review into research proposals.
- We shall offer a transparent appeals and review process if an account is restricted for alleged misuse (e.g. account suspension).
- We shall reinstate researchers' accounts in the event that of good faith research initiatives are found to have been penalized.

Good Faith Researcher Commitments

Scope of Research – Investigation into behavior of the AI system, including those disallowed by the acceptable use policy.

Researcher Responsibilities – 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:

- In-scope: Test only in-scope systems and respect out-of-scope systems.
- Vulnerability disclosure: 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).
- Harms to users and systems: Refrain from violating privacy, disrupting systems, destroying data, or harming user experience.
- Privacy requirements: 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.

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.

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 (Evtimov et al., 2019). Section 1201 of the Digital Millennium Copyright Act (DMCA) allows for civil lawsuits if researchers circumvent technological protection measures, 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 hampered 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) by alleging that New York Times research into its model constituted hacking (Brittain, 2024). Relatedly, a petition for a DMCA exemption has been filed requesting that researchers be allowed to investigate bias in generative AI systems (Weiss, 2023). The Computer Crime and Intellectual Property Section of the US Department of Justice's Criminal Division has supported the petition, calling for an exemption to DMCA § 1201 for "good-faith research into bias and other harmful and unlawful outputs" of AI models and systems (Lynch, Jr., 2024).

In the context of social media research, Abdo et al. (2022) argue a safe harbor should be 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 to enforce their policies against bad actors.

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 sys-

tems (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.[5] Recent legal victories have narrowed the applicability of CFAA with respect to "online discrimination studies" and other forms of platform accountability research, which could also protect some forms of research related to generative AI (Van Buren v. United States; Sandvig v. Barr, 2020; Sandvig v. Sessions, 2018).

We recommend that 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.[6]

[5]To be clear, there may be differences between the legal sensitivity of generative AI evaluation and research on social media platforms, in particular because of the privacy-sensitive nature of user data on social media.

[6]There are several other hindrances to an independent evaluation of generative AI systems that are important but out of scope for this work, such as the lack of details about how models are versioned or how long model versions will be available, the lack of visibility into what safety filters are implemented, and the lack of documentation of how to use models for specific use cases.

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 recommendations, grounded in prior proposals. First, we discuss how to scale up participation by delegating responsibilities to trusted independent third parties to *pre-review* researcher access. Then we discuss how an independently reviewed and transparent account suspension appeals process could enable fairer *post-review* to researcher access. Independent review and scaling participation are staples of both options.

Independent third parties can scale participation in AI evaluation, without misaligned corporate incentives. To facilitate more equitable access, and reduce the potential for corporate favoritism, we propose the responsibility of access authorization be delegated to trusted third parties, such as universities, government, or civil society organizations. The U.S. National Artificial Intelligence Research Resource (NAIRR) offers a suitable vehicle for a pilot of this approach as it already partners with and shares resources between AI developers and nonprofits. Firms provide resource credits through NAIRR, and OpenAI has called for wider participation: "by providing broader access to essential tools and data, we are opening doors for a diverse range of talents and ideas, furthering innovation and ensuring that AI development continues to be a force for the greater good" (National Science Foundation, 2024).

A similar approach has already been adopted to provide independent access to Meta's social media user data, with the University of Michigan as the trusted intermediary (González-Bailón et al., 2023). This solution scales, with partner organizations likely to aid in access review in exchange for wider participation in AI red teaming. It also effectively diverts responsibility from corporate interests to organizations already invested in fair, responsible, and accountable AI research. These partnerships do not require AI developers to fully relinquish access control but are a meaningful step in facilitating more equitable access without stretching their own resources. Each partner organization's API usage could be traced to their API keys—essentially a "researcher API". Organizations would have autonomy to authorize their own network of researchers, but would be responsible for any misuse tied to their API keys.

A number of similar proposals (see Section 5) have been made for independent researcher access, like structured access or review boards, both of which would delegate the responsibility of access selection to independent third parties. While this approach scales well and adopts independent access privileges, it can have severe limitations if firms select only 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 and nonprofits in NAIRR such as AI2, EleutherAI, and MLCommons. These changes alone 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. Providing a fair process to enable good faith research without the fear of unjustified account suspensions is key for either a pre-review or a post-review technical safe harbor. In Appendix C we sketch an implementation of a pre-registration and appeals process, based on 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, and technical safe harbors might themselves have limitations based on how access if granted (as we have seen with social media platform research APIs (Lurie, 2023)). Still, the technical safe harbor we describe is a precondition for more independent and broader participation across all these axes, should companies offer earlier, deeper, or subsidized access. While AI companies' efforts 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 called for independent evaluation and red teaming. The White House Voluntary Commitments state that firms must "Incent third-party discovery and reporting of issues and vulnerabilities... Companies making this commitment recognize that AI systems may continue to have weaknesses and vulnerabilities even after robust red-teaming. They commit to establishing for systems within scope bounty systems, contests, or prizes to incent the responsible disclosure of weaknesses, such as unsafe behaviors, or to include AI systems in their existing bug bounty program" (White House, 2023). The EU AI Act requires providers of general-purpose AI models with systemic risks to 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" as part of their technical documentation (European Council, 2024; Hacker, 2023). In addition, Canada's voluntary commitments include a commitment that developers will "conduc[t] third-party audits prior to release" (ISED Canada, 2023).

6. Conclusion

Academics, journalists, and civil society have all highlighted the need for independent AI evaluation. Examining challenges to external evaluation of generative AI systems, we identify legal and technical safe harbors as minimum fundamental protections. These measures would significantly improve norms in the ecosystem and drive more inclusive community efforts to tackle the risks of generative AI.

Impact Statement

This paper argues that the current conditions for independent evaluation of AI systems is being deterred, and that this is problematic for broader AI safety. The proposed legal and technical protections for good faith research are designed to improve long term vulnerability research. Nonetheless, it is possible (though unlikely) that they might have short-term negative effects on misuse. For instance, malicious actors might take advantage of vulnerabilities publicly shared by researchers, if the companies are unable to patch fixes within the disclosure window. Also, malicious actors might try to abuse the safe harbor protections for their own misconduct. We have discussed these possibilities in Section 4, as has prior work (Abdo et al., 2022), and believe our design should thwart most malicious efforts. We would also posit that the benefits of more protected and inclusive safety research outweigh the risks of attempted misuse, particularly given the status quo, but believe that trade-off should be carefully explored.

Implementation of these safe harbors would likely shape future interactions between AI safety researchers and generative AI companies along a few dimensions: (1) the field may experience a growth in independent or community-driven AI safety evaluation where researchers can freely access the models to safely conduct evaluation in a regulated setting, (2) the AI companies that provide research access and protections may receive more attention and vulnerability analysis than those that don't, and (3) more deeply evaluated AI systems are likely to have more mature guardrails built around them, which could reduce the likelihood of safety vulnerabilities being exploited by malicious actors.

It is also important to consider that these proposals are set within the context of US legal perspectives. We propose this as a direction of future work, detailed in Appendix A.

Acknowledgements

We thank Alex Abdo, Amit Elazari, Daphne Keller, Deb Raji, Divyansh Kaushik, Dylan Hadfield-Menell, Harley Geiger, Katy Glenn Bass, Ramya Krishnan, Riana Pfefferkorn, Shreyas Gandlur, Stephen Casper, Yacine Jernite, and Zach Lipton for their insightful feedback and guidance. We thank the staff of the Knight First Amendment Institute for their support in editing a blog post based on this paper. We also thank the 350+ researchers, journalists, and public-interest technologists who signed our open letter advocating for a safe harbor for independent AI evaluation.

References

Abdo, A., Krishnan, R., Krent, S., Welber Falcón, E., and Woods, A. K. A safe harbor for platform research. Knight Columbia, 1 2022. URL https://knightcolumbia.org/content/a-safe-harbor-for-platform-research.

Ada Lovelace Institute. Post-summit civil society communique, 11 2023. URL https://www.adalovelaceinstitute.org/news/post-summit-civil-society-communique/.

Akgul, O., Eghtesad, T., Elazari, A., Gnawali, O., Grossklags, J., Mazurek, M. L., Votipka, D., and Laszka, A. Bug hunters' perspectives on the challenges and benefits of the bug bounty ecosystem. In *32nd USENIX Security Symposium (USENIX Security). https://doi. org/10.48550/arXiv*, volume 2301, 2023.

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., Schuett, J., Shavit, Y., Siddarth, D., Trager, R., and Wolf, K. Frontier ai regulation: Managing emerging risks to public safety, 2023.

Anthropic. Core views on ai safety: When, why, what, and how. https://www.anthropic.com/news/core-views-on-ai-safety, 3 2023a.

Anthropic. Frontier threats red teaming for ai safety. Anthropic, 7 2023b. URL https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety.

Anthropic. Responsible disclosure policy, December 2023. URL https://www.anthropic.com/responsible-disclosure-policy.

Barclay, L. Facebook banned me for life because i help people use it less, 10 2021. URL https://slate.com/technology/2021/10/facebook-unfollow-everything-cease-desist.html.

Barrabi, T. Sam altman — who warned ai poses 'risk of extinction' to humanity — is also a 'doomsday prepper'. *New York Post*, 6 2023. URL https://nypost.com/2023/06/05/sam-altman-who-warned-ai-poses-risk-of-extinction-to-humanity-is-also-a-doomsday-prepper/.

Belanger, A. 100+ researchers say they stopped studying x, fearing elon musk might sue them. https://arstechnica.com/tech-policy/2023/11/100-researchers-say-they-stopped-studying-x-fearing-elon-musk-might-sue-them/, 11 2023.

Birhane, A., Steed, R., Ojewale, V., Vecchione, B., and Raji, I. D. Ai auditing: The broken bus on the road to ai accountability, 2024.

Blog, G. Rebooting responsible disclosure: a focus on protecting end users. https://security.googleblog.com/2010/07/rebooting-responsible-disclosure-focus.html, 7 2010.

Boiko, D. A., MacKnight, R., and Gomes, G. Emergent autonomous scientific research capabilities of large language models, 2023.

Bommasani, R., Klyman, K., Longpre, S., Kapoor, S., Maslej, N., Xiong, B., Zhang, D., and Liang, P. The foundation model transparency index, 2023a.

Bommasani, R., Zhang, D., Lee, T., and Liang, P. Improving transparency in ai language models: A holistic evaluation. *Foundation Model Issue Brief Series*, 2023b. URL https://hai.stanford.edu/foundation-model-issue-brief-series.

Boyd, D., DiResta, R., Donovan, J., douek, e., Frye, E., Gleicher, N., Raji, D., Rid, T., Roth, Y., Wanless, A., and Wolf, C. Commission on information disorder final report. Technical report, Aspen Institute, November 2021. URL https://www.aspeninstitute.org/wp-content/uploads/2021/11/Aspen-Institute_Commission-on-Information-Disorder_Final-Report.pdf. Recommendations for transparency.

Brittain, B. OpenAI says New York Times 'hacked' ChatGPT to build copyright lawsuit. *Reuters*, Feb 2024. URL https://www.reuters.com/technology/cybersecurity/openai-says-new-york-times-hacked-chatgpt-build-copyright-lawsuit-2024-02-27/.

Brodkin, J. Missouri threatens to sue a reporter who flagged a security flaw. https://www.wired.com/story/missouri-threatens-sue-reporter-state-website-security-flaw/, 10 2021.

Bucknall, B. S. and Trager, R. F. Structured access for third-party research on frontier ai models: Investigating researchers' model access requirements, 2023. URL https://www.governance.ai/research-paper/structured-access-for-third-party-research-on-frontier-ai-models.

Bugcrowd. Vulnerability disclosure policy: What is it & why is it important? Bugcrowd Blog, 12 2023. URL https://www.bugcrowd.com/blog/vulnerability-disclosure-policy-what-is-it-why-is-it-important/.

Burtell, M. and Woodside, T. Artificial influence: An analysis of ai-driven persuasion. *arXiv preprint arXiv:2303.08721*, 2023.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., Haupt, A., Wei, K., Scheurer, J., Hobbhahn, M., Sharkey, L., Krishna, S., Hagen, M. V., Alberti, S., Chan, A., Sun, Q., Gerovitch, M., Bau, D., Tegmark, M., Krueger, D., and Hadfield-Menell, D. Black-box access is insufficient for rigorous ai audits, 2024.

Cattell, S. and Ghosh, A. Coordinated disclosure for ai: Beyond security vulnerabilities. *arXiv preprint arXiv:2402.07039*, 2024.

CFAA. Computer Fraud and Abuse Act. 18 U.S.C. § 1030, 1986.

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., and Wong, E. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023.

Colannino, J. The copyright office expands your security research rights. https://github.blog/2021-11-23-copyright-office-expands-security-research-rights/, 23 2021.

Commission, F. T. The ftc voice cloning challenge. https://www.ftc.gov/news-events/contests/ftc-voice-cloning-challenge, 2023.

Costanza-Chock, S., Raji, I. D., and Buolamwini, J. Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1571–1583, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533213. URL https://doi.org/10.1145/3531146.3533213.

DeLong, L. A. Facebook disables ad observatory; academicians and journalists fire back. NYU Center for Cybersecurity, 8 2021. URL https://cyber.nyu.edu/2021/08/21/facebook-disables-ad-observatory-academicians-and-journalists-fire-back/.

Department of Justice. Department of justice announces new policy for charging cases under the computer fraud and abuse act. Press Release, 5 2022. URL https://www.justice.gov/opa/pr/department-justice-announces-new-policy-charging-cases-under-computer-fraud-and-abuse-act.

Deshpande, A., Murahari, V., Rajpurohit, T., Kalyan, A., and Narasimhan, K. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*, 2023.

DiResta, R., Edelson, L., Nyhan, B., and Zuckerman, E. It's time to open the black box of social media. https://www.scientificamerican.com/article/its-time-to-open-the-black-box-of-social-media/, 4 2022. 5 min read.

disclose.io. Research threats: Legal threats against security researchers. https://github.com/disclose/research-threats, 2021.

DMCA. Digital Millennium Copyright Act. 17 U.S.C. § 1201, 1998a.

DMCA. Digital millennium copyright act. 17 U.S.C. § 1204(a), 1998b.

Douglas Heaven, W. How to make a chatbot that isn't racist or sexist. MIT Technology Review, 10 2020. URL https://www.technologyreview.com/2020/10/23/1011116/chatbot-gpt3-openai-facebook-google-safety-fix-racist-sexist-language-ai/.

Elazari, A. We Need Bug Bounties for Bad Algorithms, May 2018a. URL https://www.vice.com/en/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms.

Elazari, A. Hacking the law: Are bug bounties a true safe harbor? In *Enigma 2018 (Enigma 2018)*, 2018b.

Elazari, A. Private ordering shaping cybersecurity policy: The case of bug bounties. *An edited, final version of this paper in Rewired: Cybersecurity Governance, Ryan Ellis and Vivek Mohan eds. Wiley*, 2019.

Etcovich, D. and van der Merwe, T. Coming in from the cold: A safe harbor from the cfaa and the dmca §1201 for security researchers. Berkman Klein Center Research Publication No. 2018-4. Assembly Publication Series, Berkman Klein Center for Internet & Society, Harvard University, 2018. URL http://nrs.harvard.edu/urn-3:HUL.InstRepos:37135306.

European Council. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2024. URL https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

Evtimov, I., O'Hair, D., Fernandes, E., Calo, R., and Kohno, T. Is tricking a robot hacking? *Berkeley Technology Law Journal*, 34(3):891–918, 2019.

Executive Office of the President. Safe, secure, and trustworthy development and use of artificial intelligence. Executive Order, 10 2023. URL https://www.federalregister.gov/documents/2023/10/30/2023-24110/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence. Federal Register Vol. 88, No. 210 (October 30, 2023).

Fang, R., Bindu, R., Gupta, A., Zhan, Q., and Kang, D. Llm agents can autonomously hack websites. *arXiv preprint arXiv:2402.06664*, 2024.

Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., and Chen, B. J. Ai red-teaming is not a one-stop solution to ai harms: Recommendations for using red-teaming for ai accountability. Data & Society, 10 2023. URL https://datasociety.net/library/ai-red-teaming-is-not-a-one-stop-solution-to-ai-harms-recommendations-for-using-red-teaming-for-ai-accountability/.

Ge, S., Zhou, C., Hou, R., Khabsa, M., Wang, Y.-C., Wang, Q., Han, J., and Mao, Y. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*, 2023.

Gil, A., Neelbauer, J., and A. Schweidel, D. Generative ai has an intellectual property problem. Harvard Business Review, 04 2023. URL https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., Iyengar, S., Kim, Y. M., Malhotra, N., Moehler, D., Nyhan, B., Pan, J., Rivera, C. V., Settle, J., Thorson, E., Tromble, R., Wilkins, A., Wojcieszak, M., de Jonge, C. K., Franco, A., Mason, W.,

Stroud, N. J., and Tucker, J. A. Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398, 2023. doi: 10.1126/science.ade7138. URL https://www.science.org/doi/abs/10.1126/science.ade7138.

Greene, T. C. Sdmi cracks revealed. https://www.theregister.com/2001/04/23/sdmi_cracks_revealed/, 4 2001.

Grynbaum, M. M. and Mac, R. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. *The New York Times*, Dec 2023. URL https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Gupta, R. Laion and the challenges of preventing ai-generated csam. https://www.techpolicy.press/laion-and-the-challenges-of-preventing-ai-generated-csam/, 1 2024.

Hacker, P. Comments on the final trilogue version of the ai act, January 2023. URL https://media.licdn.com/dms/document/media/D4E1FAQE9w01juCUvIw/feedshare-document-pdf-analyzed/0/1706022316786?e=1707350400&v=beta&t=PQMy2m6nOfRLfkHd4pO-ZJ0JJWvehexHNLmWJLgLYrA.

HackerOne. Hackerone gold standard safe harbor. HackerOne, 2023. URL https://hackerone.com/security/safe_harbor.

Hansen, R. and Venables, P. Introducing google's secure ai framework, June 2023. URL https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks. *arXiv preprint arXiv:2306.12001*, 2023.

Horwitz, J., Wells, G., Seetharaman, D., Hagey, K., Scheck, J., Purnell, N., Schechner, S., and Glazer, E. The facebook files: A wall street journal investigation. https://www.wsj.com/articles/the-facebook-files-11631713039, 2021.

Hu, K. Chatgpt sets record for fastest-growing user base - analyst note. *Reuters*, February 2023. URL https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Huang, Y., Gupta, S., Xia, M., Li, K., and Chen, D. Catastrophic jailbreak of open-source llms via exploiting generation, 2023a.

Huang, Y., Gupta, S., Zhong, Z., Li, K., and Chen, D. Privacy implications of retrieval-based language models. *arXiv preprint arXiv:2305.14888*, 2023b.

Inflection. Our policy on frontier safety, 2023. URL https://inflection.ai/frontier-safety.

ISED Canada. Voluntary code of conduct on the responsible development and management of advanced generative ai systems, September 2023. URL https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Jonathan, S. Ny times sues openai, microsoft for infringing copyrighted works. Reuters, 12 2023. URL https://www.reuters.com/legal/transactional/ny-times-sues-openai-microsoft-infringing-copyrighted-work-2023-12-27/.

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., et al. On the societal impact of open foundation models. 2024.

Kenway, J., François, C., Costanza-Chock, S., Raji, I. D., and Buolamwini, J. Bug bounties for algorithmic harms?, 2022. URL https://www.ajl.org/bugs.

Klyman, K. Acceptable use policies for foundation models: Considerations for policymakers and developers. Stanford Center for Research on Foundation Models, April 2024. URL https://crfm.stanford.edu/2024/04/08/aups.html.

Klyman, K., Longpre, S., Kapoor, S., Narayanan, A., Korolova, A., and Henderson, P. Comments from researchers affiliated with mit, princeton center for information technology policy, and stanford center for research on foundation models, ninth triennial proceeding, class 4. United States Copyright Office, 2024. URL https://www.regulations.gov/comment/COLC-2023-0004-0111.

Kotha, S., Springer, J. M., and Raghunathan, A. Understanding catastrophic forgetting in language models via implicit inference. *arXiv preprint arXiv:2309.10105*, 2023.

Krawiec, K. D. Cosmetic compliance and the failure of negotiated governance. *Wash. ULQ*, 81:487, 2003.

Lakatos, S. A revealing picture: Ai-generated 'undressing' images move from niche pornography discussion forums to a scaled and monetized online business. Technical report, Graphika, Dec 2023. URL https://public-assets.graphika.com/reports/graphika-report-a-revealing-picture.pdf.

Lambert, N. Undoing rlhf and the brittleness of safe llms, 10 2023. URL https://www.interconnects.ai/p/undoing-rlhf.

Li, H., Guo, D., Fan, W., Xu, M., and Song, Y. Multi-step jailbreaking privacy attacks on chatgpt. *arXiv preprint arXiv:2304.05197*, 2023a.

Li, J., Yang, Y., Wu, Z., Vydiswaran, V., and Xiao, C. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. *arXiv preprint arXiv:2304.14475*, 2023b.

Liang, P., Bommasani, R., Creel, K. A., and Reich, R. The time is now to develop community norms for the release of foundation models, 2022. URL https://crfm.stanford.edu/2022/05/17/community-norms.html.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Liu, C., Zhao, F., Qing, L., Kang, Y., Sun, C., Kuang, K., and Wu, F. Goal-oriented prompt attack and safety evaluation for llms. *arXiv e-prints*, pp. arXiv–2309, 2023.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023.

Lurie, E. Comparing platform research api requirements. Tech Policy Press, 2023. URL https://www.techpolicy.press/comparing-platform-research-api-requirements/.

Lynch, Jr., J. T. Letter from department of justice criminal division – computer crime and intellectual property section to usco re: Class 4. Computer Crime and Intellectual Property Section, Criminal Division, United States Department of Justice, 2024. URL https://www.copyright.gov/1201/2024/USCO-letters/Letter%20from%20Department%20of%20Justice%20Criminal%20Division.pdf.

Marcus, G. and Southen, R. Generative ai has a visual plagiarism problem. *IEEE Spectrum*, 1 2024. URL https://spectrum.ieee.org/midjourney-copyright.

Maus, N., Chao, P., Wong, E., and Gardner, J. R. Black box adversarial prompting for foundation models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023.

Meta. Overview of meta ai safety policies prepared for the uk ai safety summit, 2023. URL https://transparency.fb.com/en-gb/policies/ai-safety-policies-for-safety-summit/.

Midjourney. Terms of service, December 2023. URL https://docs.midjourney.com/docs/terms-of-service.

Moore, D. A., Tetlock, P. E., Tanlu, L., and Bazerman, M. H. Conflicts of interest and the case of auditor independence: Moral seduction and strategic issue cycling. *Academy of management review*, 31(1):10–29, 2006.

Mozilla. How safe are our online platforms? let's open the door for social media researchers. https://foundation.mozilla.org/en/campaigns/unknown-influence/, 2023.

Narayanan, A. and Kapoor, S. Model alignment protects against accidental harms, not intentional ones, 12 2023a. URL https://www.aisnakeoil.com/p/model-alignment-protects-against.

Narayanan, A. and Kapoor, S. Generative ai companies must publish transparency reports, 2023b. URL https://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tramèr, F., and Lee, K. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

National Science Foundation. Democratizing the future of ai r&d: Nsf to launch national ai research resource pilot. https://new.nsf.gov/news/democratizing-future-ai-rd-nsf-launch-national-ai, 1 2024.

Nelson, C. and Rose, S. https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio, 10 2023. URL https://www.longtermresilience.org/post/report-launch-examining-risks-at-the-intersection-of-ai-and-bio.

NIST. Nist seeks collaborators for consortium supporting artificial intelligence safety, 2023. URL https://www.nist.gov/news-events/news/2023/11/nist-seeks-collaborators-consortium-supporting-artificial-intelligence.

NIST. Test, evaluation & red-teaming, 2024. URL https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence/test.

OpenAI. Introducing chatgpt and whisper apis. 2023a. URL https://openai.com/blog/introducing-chatgpt-and-whisper-apis.

OpenAI. Sharing and publication policy. https://openai.com/policies/sharing-publication-policy#research, 2023b.

OpenAI. Bugcrowd: Openai, 2024a. URL https://bugcrowd.com/openai.

OpenAI. Researcher access program application, 2024b. URL https://openai.com/form/researcher-access-program.

Pa Pa, Y. M., Tanizaki, S., Kou, T., Van Eeten, M., Yoshioka, K., and Matsumoto, T. An attacker's dream? exploring the capabilities of chatgpt for developing malware. In *Proceedings of the 16th Cyber Security Experimentation and Test Workshop*, pp. 10–18, 2023.

Park, J., Singh, V., and Wisniewski, P. Supporting youth mental and sexual health information seeking in the era of artificial intelligence (ai) based conversational agents: Current landscape and future directions. *Available at SSRN 4601555*, 2023.

Parrish, A., Kirk, H. R., Quaye, J., Rastogi, C., Bartolo, M., Inel, O., Ciro, J., Mosquera, R., Howard, A., Cukierski, W., et al. Adversarial nibbler: A data-centric challenge for improving the safety of text-to-image models. *arXiv preprint arXiv:2305.14384*, 2023.

Persily, N. A proposal for researcher access to platform data: The platform transparency and accountability act. *Journal of Online Trust and Safety*, 1(1), 2021.

Pfefferkorn, R. America's anti-hacking laws pose a risk to national security. https://www.brookings.edu/articles/americas-anti-hacking-laws-pose-a-risk-to-national-security/, 9 2021.

Pfefferkorn, R. Shooting the messenger: Remediation of disclosed vulnerabilities as cfaa "loss". *Richmond Journal of Law & Technology*, 29:89, 2022. URL https://jolt.richmond.edu/files/2022/11/Pfefferkorn-Manuscript-Final.pdf.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., and Zhang, Y. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. *arXiv preprint arXiv:2305.13873*, 2023.

Raji, I. D., Xu, P., Honigsberg, C., and Ho, D. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pp. 557–571, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534181. URL https://doi.org/10.1145/3514094.3534181.

Rando, J., Paleka, D., Lindner, D., Heim, L., and Tramèr, F. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Renaud, K., Warkentin, M., and Westerman, G. *From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI*. MIT Sloan Management Review, 2023.

Robey, A., Wong, E., Hassani, H., and Pappas, G. J. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*, 2023.

Sandvig v. Barr. United states district court for the district of columbia. 451 F. Supp. 3d 73 (D.D.C. 2020), 2020.

Sandvig v. Sessions. United states district court for the district of columbia. 315 F. Supp. 3d 1 (D.D.C. 2018), 2018.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? *arXiv preprint arXiv:2303.17548*, 2023.

Shah, R., Montixi, Q. F., Pour, S., Tagade, A., and Rando, J. Scalable and transferable black-box jailbreaks for language models via persona modulation. In *Socially Responsible Language Modelling Research*, 2023.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*, 2023.

Shen, X., Chen, Z., Backes, M., Shen, Y., and Zhang, Y. " do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models. *arXiv preprint arXiv:2308.03825*, 2023.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *ICLR*, 2024.

Soice, E. H., Rocha, R., Cordova, K., Specter, M., and Esvelt, K. M. Can large language models democratize access to dual-use biotechnology? *arXiv preprint arXiv:2306.03809*, 2023.

Solaiman, I., Talat, Z., Agnew, W., Ahmad, L., Baker, D., Blodgett, S. L., au2, H. D. I., Dodge, J., Evans, E., Hooker, S., Jernite, Y., Luccioni, A. S., Lusoli, A., Mitchell, M., Newman, J., Png, M.-T., Strait, A., and Vassilev, A. Evaluating the social impact of generative ai systems in systems and society, 2023.

Stupp, C. Fraudsters used ai to mimic ceo's voice in unusual cybercrime case. https://www.wsj.com/articles/fraudsters-used-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567098001, 8 2019. WSJ PRO.

Suleyman, M. and Bhaskar, M. *The Coming Wave: Technology, Power, and the Twenty-First Century's Greatest Dilemma*. Penguin Random House, 2023.

Sven Cattell. Generative red team recap, Oct 2023. URL https://aivillage.org/defcon%2031/generative-recap/.

Tabassi, E. Artificial intelligence risk management framework (ai rmf 1.0), 2023-01-26 05:01:00 2023. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

The Hacking Policy Council, Dec 2023. URL https://assets-global.website-files.com/62713397a014368302d4ddf5/6579fcd1b821fdc1e507a6d0_Hacking-Policy-Council-statement-on-AI-red-teaming-protections-20231212.pdf.

Thiel, D., Stroebel, M., and Portnoff, R. Generative ml and csam: Implications and mitigations, 2023. URL https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations.

Toner, H. and McCauley, T. Ai firms mustn't govern themselves, say ex-members of openai's board. The Economist, 2024. URL https://www.economist.com/by-invitation/2024/05/26/ai-firms-mustnt-govern-themselves-say-ex-members-of-openais-board.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Urbina, F., Lentzos, F., Invernizzi, C., and Ekins, S. Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, 4(3):189–191, 2022.

Van Buren v. United States. Supreme court of the united states. No. 19-783, 141 S.Ct. 1648. 18 U.S.C. §1030.

Walshe, T. and Simpson, A. Towards a greater understanding of coordinated vulnerability disclosure policy documents. *Digital Threats: Research and Practice*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2023.

Weidinger, L., Rauh, M., Marchal, N., Manzini, A., Hendricks, L. A., Mateos-Garcia, J., Bergman, S., Kay, J., Griffin, C., Bariach, B., Gabriel, I., Rieser, V., and Isaac, W. S. Sociotechnical safety evaluation of generative ai systems. *ArXiv*, abs/2310.11986, 2023. URL https://api.semanticscholar.org/CorpusID:264289156.

Weiss, J. Petition for new exemption to section 1201 of the digital millenium copyright act: Exemption for security research pertaining to generative ai bias, June 2023. URL https://www.copyright.gov/1201/2024/petitions/proposed/New-Pet-Jonathan-Weiss.pdf.

White House. Ensuring safe, secure, and trustworthy ai, 2023. URL https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

Whittaker, M. The steep cost of capture. *Interactions*, 28 (6):50–55, 2021.

Xiang, C. 'he would still be here': Man dies by suicide after talking with ai chatbot, widow says. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says, 3 2023.

Xu, R., Lin, B. S., Yang, S., Zhang, T., Shi, W., Zhang, T., Fang, Z., Xu, W., and Qiu, H. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. *arXiv preprint arXiv:2312.09085*, 2023.

Yang, X., Wang, X., Zhang, Q., Petzold, L., Wang, W. Y., Zhao, X., and Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949*, 2023.

Yong, Z. X., Menghini, C., and Bach, S. Low-resource languages jailbreak gpt-4. In *Socially Responsible Language Modelling Research*, 2023.

Yu, J., Lin, X., and Xing, X. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts. *arXiv preprint arXiv:2309.10253*, 2023.

Zalnieriute, M. "transparency-washing" in the digital age : A corporate agenda of procedural fetishism. Technical report, 2021. URL http://hdl.handle.net/11159/468588.

Zeng, Y., Lin, H., Zhang, J., Yang, D., Jia, R., and Shi, W. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*, 2024.

Zhan, Q., Fang, R., Bindu, R., Gupta, A., Hashimoto, T., and Kang, D. Removing RLHF Protections in GPT-4 via Fine-Tuning. *arXiv preprint arXiv:2311.05553*, 2023.

Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

Zhao, X., Yang, X., Pang, T., Du, C., Li, L., Wang, Y.-X., and Wang, W. Y. Weak-to-strong jailbreaking on large language models, 2024.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

Zuboff, S. The age of surveillance capitalism. In *Social Theory Re-Wired*, pp. 203–213. Routledge, 2023.

AI Company	AI System	Acceptable Use Policy	Deep Access	Researcher Access	Bug Bounty	Safe Harbor	Enforcement Process	Enforcement Justification	Enforcement Appeal
OpenAI	GPT-4	●[Link]	◑[Link]	●[Link]	●[Link]	◑[Link]	●[Link]	○[Link]	◑[Link]
Google	Gemini 1.0	●[Link]	○	○	●[Link]	○[Link]	○[Link]	◑[Link]	○[Link]
Anthropic	Claude 2	●[Link]	○	◑[Link]	○	◑[Link]	●[Link]	○[Link]	○[Link]
Inflection	Inflection-1	●[Link]	○	○	○	○	○[Link]	◑[Link]	◑[Link]
Meta	Llama 2	●[Link]	●[Link]	●[Link]	●[Link]	◑[Link]	○[Link]	○[Link]	○[Link]
Midjourney	Midjourney v6	●[Link]	○	○	○	○	○	○	◑[Link]
Cohere	Command	●[Link]	○	●[Link]	○	◑[Link]	○[Link]	○[Link]	○[Link]

Table A1. **A summary of the policies, access, and enforcement for major AI systems, with links to evidence where applicable.** ● indicates that a company satisfies or provides access to information in a column, ○ indicates it does not, and ◑ indicates partial satisfaction.

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.

- **Acceptable Use Policy:** ● indicates that the company documents its acceptable use 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

only 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 acceptable use 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 acceptable use 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 acceptable use 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 GPT-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 or 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 acceptable use 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 decommisioned), 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 acceptable use 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 acceptable use 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 acceptable use 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 company 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.[7] 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.

[7] We emailed "papers@openai.com" to ask for clarification on research exemptions for OpenAI's Usage Policies, but received no response.