Stanford University
Human-Centered
Artificial Intelligence

Policy Brief
HAI Policy & Society
February 2025

Safeguarding Third-Party AI Research

Kevin Klyman, Shayne Longpre, Sayash Kapoor,
Rishi Bommasani, Percy Liang, Peter Henderson

THIRD-PARTY EVALUATION IS A CORNERSTONE OF EFFORTS TO REDUCE THE SUBSTANTIAL RISKS POSED BY AI SYSTEMS. AI is a vast field with thousands of highly specialized experts around the world who can help stress-test the most powerful systems. But few companies empower these researchers to test their AI systems, for fear of exposing flaws in their products. AI companies often block safety research with restrictive terms of service or by suspending researchers who report flaws.

In our paper, "A Safe Harbor for AI Evaluation and Red Teaming," we assess the policies and practices of seven top developers of generative AI systems, finding that none offers comprehensive protections for third-party AI research. Unlike with cybersecurity, generative AI is a new field without well-established norms regarding flaw disclosure, safety standards, or mechanisms for conducting third-party research. We propose that developers adopt safe harbors to enable good-faith, adversarial testing of AI systems.

Key Takeaways

Third-party AI research is essential to ensure that AI companies do not grade their own homework, but few companies actively protect or promote such research.

We found no major foundation model developers currently offer comprehensive protections for third-party evaluation. Instead, their policies often disincentivize it.

A safe harbor for good-faith research should be a top priority for policymakers. It enables good-faith research and increases the scale, diversity, and independence of evaluations.

Introduction

Generative AI systems pose a wide range of potential risks, from enabling the creation of nonconsensual intimate imagery to facilitating the development of malware. Evaluating generative AI systems is crucial to understanding the technology, ensuring public accountability, and reducing these risks.

In July 2023, many prominent AI companies signed voluntary commitments at the White House, pledging to "incent third-party discovery and reporting of issues and vulnerabilities." More than a year later, implementation of this commitment has been uneven. While some companies do reward researchers for finding security flaws in their AI systems, few companies strongly encourage research on safety or provide concrete protections for good-faith research practices. Instead, leading generative AI companies' terms of service legally prohibit third-party safety and trustworthiness research, in effect threatening anyone who conducts such research with bans from their platforms or even legal action. For example, companies' policies do not allow researchers to jailbreak AI systems like ChatGPT, Claude, or Gemini to assess potential threats to U.S. national security.

In March 2024, we penned an open letter signed by over 350 leading AI researchers and advocates calling for a safe harbor for third-party AI evaluation. The researchers noted that while security research on traditional software is protected by voluntary company protections (safe harbors), established vulnerability disclosure norms, and legal safeguards from the Department of Justice, AI safety and trustworthiness research lacks comparable protections.

We assess the policies and practices of seven top developers of generative AI systems, finding that none offers comprehensive protections for third-party AI research.

Companies have continued to be opaque about key aspects of their most powerful AI systems, such as the data used to build their models. Developers of generative AI models tout the safety of their systems based on internal red teaming, but there is no way for the government or independent researchers to validate these results, as companies do not release reproducible evaluations.

Generative AI companies also impose barriers on their platforms that limit good-faith research. Similar issues plague social media: Companies have taken steps to prevent researchers and journalists from conducting investigations on their platforms that, together with federal legislation, have had a chilling effect on such research and worsened the spread of harmful content online. But conducting research on generative AI systems comes with additional challenges, as the content on generative AI platforms is not publicly available. Users need accounts to access AI-generated content, which can be restricted by the company that owns the platform. Many AI companies also block

certain user requests and limit the functionality of their models to prevent researchers from unearthing issues related to safety or trustworthiness. The stakes are also higher for AI, which has the potential not only to turbocharge misinformation but also to provide U.S. adversaries like China and Russia with material strategic advantages.

To assess the state of independent evaluation for generative AI, our team of machine learning, law, and policy experts conducted a thorough review of seven major AI companies' policies, access provisions, and related enforcement processes. We detail our experiences with evaluation of AI systems and potential barriers other third-party evaluators may face, and propose alternative practices and policies to enable broader community participation in AI evaluation.

Research Outcomes

At the time of our study, we found that AI companies' policies and practices regarding independent evaluation were inconsistent, opaque, and burdensome, as shown in Figure 1. Gaps in companies' practices force well-intentioned researchers to either wait for approval from unresponsive researcher access programs or risk violating company policy and losing access to their accounts in the course of conducting good-faith research.

Figure 1: What access protections do AI companies provide for independent safety research?

Source: Longpre et al., A Safe Harbor for AI Evaluation and Red Teaming

Company Practices	Claude 2 (Anthropic)	Command (Cohere)	Gemini (Google)	Inflection-1 (Inflection)	Llama 2 (Meta)	Midjourney v6 (Midjourney)	GPT-4 (OpenAI)
Model Access How can researchers access the company's foundation model?							
Public API	○	●	●	○	●	○	●
Deep Access	○	○	○	○	●	○	◐
Dedicated Researcher Access	◐	●	○	○	●	○	●
Independent Access Review	○	○	○	○	●	○	○
Bug Bounty	○	○	●	○	●	○	●
Safe Harbor What types of research do companies legally protect, and are those protections determined at their sole discretion?							
Security	●	○	○	○	●	○	●
AI Safety & Flaws	○	◐	○	○	○	○	◐
Not Sole Discretion	○	○	○	○	○	○	●
Policy Enforcement Transparency & Fairness Are the policies used to enforce the terms of use transparent and fair, providing violation justifications and appeals?							
Enforcement Policy	●	○	○	○	○	○	●
Enforcement Justifications	○	○	◐	◐	○	○	○
Enforcement Violation Appeals	○	○	○	◐	○	◐	◐

AI companies' policies and practices regarding independent evaluation were inconsistent, opaque, and burdensome.

While some companies provide access to their systems through public Application Programming Interfaces, allow for deeper access to their systems through fine-tuning or open-weight releases, and maintain dedicated researcher access programs, others take none of these steps. Moreover, companies generally select the researchers who participate in their researcher access programs, increasing the likelihood of favoritism and imbalanced representation in such programs.

Existing protections apply almost exclusively to security research. Developers of AI systems have engaged to differing degrees with external red teamers and evaluators, though they focus on cybersecurity-related flaws. For example, OpenAI, Google, and Meta have bug bounties (monetary rewards for individuals who report security vulnerabilities) as well as legal protections for security research. Still, companies like Meta "reserve final and sole discretion for whether you are acting in good faith and in accordance with this Policy," which could deter good-faith research. These legal protections extend only to traditional security issues like unauthorized account access and do not include broader AI safety and trustworthiness research, such as whether an AI system may facilitate the development of cyber weapons.

Cohere, OpenAI, and Anthropic are exceptions. They offer some protections beyond security research, though some ambiguity remains as to the scope of protected activities. Cohere, for example, allows "intentional stress testing of the API and adversarial attacks" provided appropriate vulnerability disclosure (without explicit legal promises). OpenAI expanded its safe harbor to include "model issues research" and in some limited cases "academic research related to model safety" in response to an early draft of our paper. Anthropic created a "model safety bug bounty" program after the release of our paper that rewards researchers for "novel, universal jailbreak attacks that could expose vulnerabilities in critical, high risk domains such as CBRN (chemical, biological, radiological, and nuclear) and cybersecurity."

Restrictions on researchers' accounts can have a chilling effect by making researchers reluctant to investigate certain risks or systems for fear of blowback.

Our experiences as researchers conducting AI evaluation surfaced common themes regarding the barriers that company practices pose to independent evaluation. For instance, restrictions on researchers' accounts can have a chilling effect by making researchers reluctant to investigate certain risks or systems for fear of blowback. Similarly, a lack of clarity regarding when and how independent researchers should disclose flaws in companies' AI systems may deter researchers from disclosing their findings, preventing safety improvements that would benefit all users of a system.

If AI companies do not voluntarily adopt such safe harbors, policymakers should mandate them.

Policy Discussion

Ensuring a safe harbor for good-faith AI safety and trustworthiness research should be a top priority for policymakers. Such protections would unlock good-faith research, increasing the scale, diversity, and independence of evaluations. It is an essential precondition for improving our understanding of the wide-ranging risks of AI, increasing public accountability, and ensuring safe and trustworthy AI.

We suggest that AI companies adopt safe harbors to improve participation, access, and incentives for public interest research into AI safety. If AI companies do not voluntarily adopt such safe harbors, policymakers should mandate them. For example, in the European Union, policymakers have already mandated protections for independent research on social media platforms.

First, companies can provide a **legal safe harbor** that protects good-faith evaluation provided it is conducted in accordance with well-established security vulnerability disclosure practices. This would allow independent researchers to stress-test important AI systems without fear of legal reprisal.

A legal safe harbor could provide assurances that AI companies will not sue researchers if their actions were taken for research purposes. In the U.S. legal regime, this would impact companies' use of the Computer Fraud and Abuse Act (CFAA) and Section 1201 of the Digital Millennium Copyright Act (DMCA) to block independent research. These risks are not theoretical; security researchers have been targeted under the CFAA, and DMCA Section 1201 hampered security research to the extent that researchers requested and won an exemption from the law for this purpose. In the context of generative AI, OpenAI has attempted to dismiss a lawsuit brought by the *New York Times* by alleging that the media company's research regarding ChatGPT when preparing the lawsuit constituted hacking.

Second, AI companies can provide a **technical safe harbor** by protecting vetted researchers from restrictions on their accounts. Such protections would prevent account suspensions or other technical

enforcement actions such as rate limiting that also impede independent safety evaluations. Each of these safe harbors should be scoped to include research activities that uncover *any* flaw in a general-purpose AI system, including all types of responses from a system that are prohibited by a company's terms of service or acceptable use policy that would not otherwise violate the law.

We propose that companies offer some path to eliminate these technical barriers for good-faith research, even when it can unearth flaws with companies' systems. This would allow for a wider variety of researchers to access AI systems and guarantee that safety research will not be foreclosed when adhering to companies' disclosure policies. One way to do this is to scale up researcher access programs and provide impartial review of applications for these programs. One potential challenge with implementing a technical safe harbor is distinguishing between legitimate research and malicious actors who ignore company policies with the intent to cause harm. An exemption to strict enforcement of companies' policies may need to be reviewed in advance, or at least when an unfair account suspension occurs. With help from credible third-party reviewers, companies can scale this review process to meet the needs of the wider research community.

As others have argued, a legal and technical safe harbor would not inhibit existing enforcement against malicious misuse, as protections are contingent on abiding by the law and stringent vulnerability disclosure policies, determined ex post. A legal safe harbor would safeguard certain research from some amount of legal liability, mitigating the deterrent of strict terms of service. Some research organizations conduct legal or institutional reviews that might otherwise be deterred by the terms of service without a carveout. A technical safe harbor would limit practical barriers to safety research from companies' enforcement of their terms of service by clarifying that researchers will not be penalized.

These protections apply only to researchers who abide by companies' vulnerability disclosure policies, to the extent researchers can subsequently justify their actions in court. Research that is already illegal or does not take reasonable steps for responsible disclosure would not succeed in claiming those protections in an ex-post investigation.

The need for independent evaluation of powerful AI systems has garnered significant support from academics, journalists, and civil society. Safe harbors would improve safety, security, and trustworthiness in the AI ecosystem, and enable community participation in urgent efforts to tackle the risks of AI.

Reference: The original article is accessible at Shayne Longpre, Sayash Kapoor, Kevin Klyman et al. 2024. Position: a safe harbor for AI evaluation and red teaming. In Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 1327, 32691–32710. https://dl.acm.org/doi/10.5555/3692070.3693397.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices.
The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Kevin Klyman is an MA candidate in international policy at Stanford University and a graduate fellow at the Stanford Center for Research on Foundation Models (CRFM), part of the Stanford Institute for Human-Centered Artificial Intelligence (HAI). His work on this brief was completed prior to becoming a government employee.

Shayne Longpre is a PhD candidate at MIT's Media Lab and the lead of the Data Provenance Initiative.

Sayash Kapoor is a senior fellow at Mozilla, a Laurance S. Rockefeller Graduate Prize Fellow in the University Center for Human Values, and a PhD candidate in computer science at Princeton University's Center for Information Technology Policy.

Rishi Bommasani is a PhD candidate in computer science at Stanford University and the society lead at Stanford CRFM.

Percy Liang is an associate professor of computer science and statistics at Stanford University and the director of Stanford CRFM.

Peter Henderson is an assistant professor of computer science and of public and international affairs at Princeton University.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: 353 Jane Stanford Way, Stanford CA 94305-5008

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**