Print Edition

Tuesday 4 August 2026

EDITION NO. 12 · 2 PIECES · 10 PAGES

IN THIS EDITION

Why smarter AI models could drive up compute prices 10x

BY DWARKESH PATEL · DWARKESH PODCAST · 3 AUGUST 2026

This is a video recording of a post I wrote last week. If you want to read the original you can check it out here.

Thanks to Mercury for sponsoring this video. Mercury's built-in AI, Command, helps me close my books and saves me a bunch of time. At the end of each month, Command categorizes my transactions and provides its rationale for every choice: I just review, fix anything that's off, and approve... and then Mercury syncs everything to QuickBooks. Get started at mercury.com/command

AI Jailbreak Disclosure Is Broken. Here's How to Fix It

BY AI FRONTIERS · AI FRONTIERS · 3 AUGUST 2026

Rich Barton-Cooper, Research Manager at MATS and **Adam Gleave**, CEO of FAR.AI — August 3, 2026

Frontier AI developers deploy significant safeguards to prevent their AI models from being misused. However, these safeguards are not robust: AI researchers consistently find techniques for circumventing them, commonly known as "jailbreaks." Once companies learn about a jailbreak, they can usually implement a fix; yet AI researchers currently lack a safe and reliable way to inform AI companies about the jailbreaks they've discovered. We propose concrete improvements to the current system based on cybersecurity norms of responsible disclosure.

Addressing this problem is urgent. Frontier AI models are already being abused by malicious actors: terrorist group Boko Haram is reported to be training its members on jailbreaking techniques in order to troubleshoot weapons, design new explosive devices, and plan attacks. Some former Boko Haram members stated an unequivocal willingness to use chemical or biological weapons—a task jailbroken models may already be able to help with today. With model capabilities continuing to rapidly accelerate —including in offensive cyber, where jailbreaks have already been employed by Russian cybercriminals—the stakes for AI misuse are escalating with every model release.

Universal jailbreaks affect even the most advanced models. Of particular concern are universal jailbreaks—jailbreaks which consistently break through safeguards across a wide variety of harmful queries. These have been found even in the most advanced models, the most recent of which elicited cyberattack assistance from OpenAI's flagship model, GPT-5.6 Sol. Research by one of the authors found tens to hundreds of universal jailbreaks in two out of four frontier models using combinations of simple, publicly available techniques. More limited jailbreaks aimed at high-risk domains are still dangerous: an alleged narrow cyber jailbreak led to the US Government shutting down access to Claude Fable 5 just days after launch.

External researchers need effective reporting channels. It is therefore a critical matter of public safety that model safeguards are tested extensively both pre- and post-deployment. In order to find and fix safeguard vulnerabilities at scale, frontier labs must provide effective channels to allow external researchers to contribute. Any jailbreak must be able to be reported quickly and safely to its developer and fixed as a matter of urgency. Additionally, the existence of jailbreaks should be disclosed to poli-

cymakers, as well as users of these systems and the wider public when it is safe to do so.

Yet currently there is often no way for third-parties to report jailbreaks to developers; most reporting mechanisms that do exist require overly broad NDAs; and developers self-grade submissions according to opaque rubrics.

We propose a two-tiered solution to address these deficiencies, drawing on established disclosure practices in cybersecurity. As soon as possible, developers should implement public reporting mechanisms for jailbreaks and share a rubric for how they evaluate jailbreak severity. In the medium term, we propose establishing a third-party clearinghouse to take in jailbreak submissions, evaluate them, disseminate the vulnerability to affected parties, and coordinate reasonable disclosure for the researcher following a fix. Together, these mechanisms will allow jailbreaks to be reported to developers for them to fix while preserving transparency.

There Is No Good Way to Report Jailbreaks

Sadly, the ecosystem for reporting jailbreak vulnerabilities is poorly set up for the task today. There are three main problems with the current regime. First, many vendors offer no official way to report jailbreaks. Second, where vendors do offer programs, the programs almost universally involve NDAs so restrictive they indefinitely prevent disclosure of the vulnerability to governments or other developers. Finally, the developers self-grade the severity of jailbreaks using opaque rubrics that are inconsistent between developers—a top-severity vulnerability with one developer might barely raise an eyebrow with another.

Independent researchers face tangled NDAs and legal risk. This means that independent researchers who find jailbreaks and wish to report them responsibly find themselves in an unfavorable position, as author Rich Barton-Cooper has personally experienced. They will be faced with inconsistent disclosure channels across frontier labs, tangled NDAs which suppress public disclosure of vulnerabilities, and opaque internal evaluation of their work by labs incentivized to downplay the severity of their findings. Additionally, researchers may not feel safe to disclose at all: probing for harmful outputs violates developers' usage policies, and safe-harbor protections may not clearly apply since jailbreaks are often explicitly excluded from standard security channels.

Institutional standing offers little advantage. Even established organizations are hampered by the status quo. One of the authors, Adam Gleave, leads FAR.AI, an AI safety non-profit that conducts pre-deployment testing for many frontier model developers and has numerous personal contacts with other developers. Even with this institutional standing, the experience is uneven: labs listen, but contest severity, set safeguard update timelines unilaterally, and provide no guarantee that they will act on a given finding.

Current Reporting Mechanisms Are Limited

OpenAI and Anthropic are the only frontier model vendors to provide any concrete jailbreak reporting routes. For chemical, biological, radiological, and nuclear (CBRN) weapon-related outputs, both run bug-bounty programs (OpenAI, Anthropic). However, a researcher must be accepted, sign an NDA, and submit into a channel that grades against an unseen rubric and permits no external disclosure without express consent.

Outside bounty programs, there is no safe disclosure method. A researcher not accepted onto a bounty program, or unwilling to be permanently silenced, has nowhere to safely disclose a CBRN jailbreak directly to developers. A powerful jailbreak is not safe to disclose openly, for example via social media, because bad actors could use it to harvest dangerous information from affected models before it is patched. Developers also frequently do not provide a clear means to route a jailbreak to the right teams; in our experience, personal outreach to contacts working on such teams has been necessary to draw attention to findings, which is not an option accessible to the wider research community.

These problems play out differently across harm domains: Anthropic runs a locked-down program for CBRN but an open, disclosure-friendly channel for cyber. What separates them appears not to be the severity of the risk, but the strength of external pressure. When Fable 5's cyber capabilities reportedly prompted the US government to suspend access to the model within days of launch, Anthropic was acutely incentivized to demonstrate a credible cyber disclosure process and has recently established an uncompensated Cyber Jailbreak disclosure program for jailbreaks targeting Fable 5. This program is markedly more transparent than usual: anyone can submit without an NDA, findings may be publicly disclosed once timing is coordinated with Anthropic, and researchers are explicitly free to report the same jailbreak to other affected vendors. But that channel is narrow: we would be excited to see Anthropic launch an analogous channel for CBRN and other jailbreak categories, and for other developers to follow suit.

Most labs exclude jailbreaks from disclosure programs entirely. The situation is much worse elsewhere: other labs, including Google DeepMind, SpaceXAI (formerly xAI), and Meta, explicitly exclude jailbreaks and model content issues from the scope of vulnerability disclosure programs, if they exist at all. Instead these programs focus on more-traditional cybersecurity issues, such as company data exfiltration, accessing other users' accounts, or authentication flaws. The Future of Life Institute (FLI), an independent non-profit that has graded frontier developers' safety practices since 2024 through a panel of external AI and governance experts, recently released their widely cited Summer 2026 Safety Index which explored each company's approach to responsible disclosure and bug bounties. The FLI judged that the leading labs—OpenAI, Anthropic, and Google DeepMind—all score a C+ in the Risk Assessment category, which translates to "uneven validity or elicitation" and "little external input." Meta and SpaceXAI score D+ and D- respectively.

Google's in-product reporting system is flawed. Google has stated they don't believe that disclosure programs for jailbreaks are the right solution, given fixing such issues "requires long-term, cross-disciplinary efforts" and instead choose to rely on reporting mechanisms built into their product. (We contest the claim that these fixes require long-term effort, as some methods enable defenders to block a whole class of jailbreaks with a few examples.) We tried out these mechanisms. The first is a "thumbs-down" of the response in which you can choose to label the content as "Offensive/Unsafe", or "Took a harmful action." There is no acknowledgement of receipt, and no means to track that action will be taken. Another in-product option is to report a legal issue, explaining why the content was unlawful in the user's country. In this case, they explicitly state that "completing and submitting this form does not guarantee that any action will be taken." Neither of these channels is appropriate for disclosing high-severity, highly detailed universal CBRN content: there is no way to describe the jailbreak method rather than specific model responses, and no guarantee of response from the vendor.

At SpaceXAI, the story is similar: their bounty program states that "model issues are out of scope for this program and should be reported through safety@x.ai." In October 2025, Barton-Cooper reported a universal jailbreak through this email address containing screenshots of several egregiously harmful model responses—including redacted evidence of clear, step-by-step guidance on building a massively destructive chemical weapon—and did not receive anything other than an automated response in return.

Fear of Liability May Incentivize Vendors to Ignore Reports

This silence may reflect more than under-resourced safety teams or disorganization. As far back as 2024, legal scholars have warned that tort liability fears can deter a developer from documenting the risks its models pose, since "creating such a record might later help plaintiffs establish negligence." A lab therefore has a perverse incentive to leave a jailbreak report unacknowledged. Liability is demonstrably a concern: in April 2026, OpenAI testified in support of SB 3444, a bill which seeks to limit liability for mass harm traceable to frontier AI. Following public backlash and Anthropic lobbying against the bill from the start—with a spokesperson calling it a "get-out-of-jail-free card against all liability"—OpenAI later walked back their support. In a world where labs are disincentivized to accept unsolicited jailbreak reports due to liability concerns, users who find effective jailbreaks are forced into bug bounty programs where they can be effectively silenced by NDA.

Established Channels Are Often Covered by Powerful NDAs

Although we applaud OpenAI and Anthropic for soliciting model testing through their bug bounty programs, we believe these channels are insufficient as a reporting mechanism. Both programs are covered by NDAs that restrict all findings submitted through these channels. Jailbreakers

are left with two undesirable options: either indefinitely lock up their findings, or—as many jailbreakers choose—publish their findings on X.

NDAs erode hard-won disclosure norms. This is not a failure unique to AI: in cybersecurity, legal scholars have documented how the recent proliferation of NDAs in bug bounties has begun to erode hard-won coordinated disclosure norms—a move the security research community has criticized. Katie Moussouris, who built Microsoft's first bug bounty and helped establish coordinated disclosure as an industry norm, says: "Why would anyone ever sign an NDA for the privilege of telling an organization what's wrong with them, especially when they may not get paid for their work?" It appears that AI labs have adopted this diminished ecosystem of indefinite legally binding prohibitions of disclosure, rather than building on the better coordinated disclosure protocols of the past. Under previous norms, researchers entered into a limited embargo period of a few months in order to give the organization time to fix the vulnerability before publishing. Findings could still be shared privately with affected parties during this time.

Current NDAs protect lab reputations more than the public. NDAs help address a serious risk: if a researcher discloses jailbreaks to the public before the company has time to fix them, then a bad actor may use those jailbreaks to cause serious harm. Yet current NDAs appear to go beyond sensible handling of such risks, and in practice do more to safeguard the reputation of individual labs than to protect the public from harm. Anthropic's program bars participants from disclosing "any jailbreaks/vulnerabilities (even resolved ones) outside of the Program without express consent" and OpenAI similarly states that "All prompts, completions, findings, and communications are covered by NDA." Additionally, cross-lab disclosure is legally ambiguous under these NDAs: a universal jailbreak which transfers across vendors does not seem to be clearly permitted to be disclosed beyond the first vendor. This means that the most potent cross-model jailbreaks submitted under bug bounty programs cannot automatically be disseminated across all affected parties today. The net effect is that the true extent of misuse risks and how quickly vendors are responding to them remain opaque to both policymakers and the public.

Labs Grade Jailbreaks Without Oversight

To make matters worse, jailbreak severity is graded by the labs themselves, often following non-public rubrics. This inevitably leads to uneven safeguards between developers, creating an inconsistent patchwork of protections. Furthermore, developers are incentivized to understate severity: developers like to boast of their models withstanding third-party red-teaming, and so acknowledging a successful universal jailbreak submission is reputationally harmful in a world increasingly waking up to misuse risk. Together the uneven landscape and incentives to downplay issues obscure the true picture of jailbreak risks.

A shared severity framework is beginning to emerge. Clarity can only come through a standardized assessment framework and timely publication of jailbreak incidents. Anthropic—in collaboration with Amazon,

Microsoft, Google, and Project Glasswing partners—recently published a Cyber Jailbreak Severity (CJS) framework following governmental intervention with Fable 5. We support this initiative, and call for other frontier vendors to coordinate on this work, both in cyber and other catastrophic risk domains.

The System Today Looks Increasingly Fragile

Anthropic has previously written that the absence of legitimate, well-compensated disclosure channels is itself a risk: without them, a malicious researcher may have an incentive to sell a jailbreak on a black market rather than informing the affected lab. Dissatisfied researchers are a second failure mode: when disclosure mechanisms break down or are not fit for purpose, some will publish vulnerabilities publicly anyway, whether to gain attention, in retaliation, or to force a rapid fix. The recent public disclosure of six Microsoft vulnerabilities by an independent cybersecurity researcher is a stark reminder of what happens when vendors overreach with legal threats or renege on good-faith commitments. There are perhaps early signs of this dynamic in the jailbreaking community: Hacker News reactions to the recent GPT-5.5 Bio Bug Bounty were largely critical, with users objecting that the NDA silences participants, that signed participants have no recourse if their submission is rejected, and that everyone outside the program has no responsible route to disclose. Capable red-teamers may already be opting out; as one user posted: "This is very much within my areas of interest, but signing an NDA in this area is a lot to ask."

Proposals for Better Disclosure Practices

We propose two complementary strategies for improving the external jailbreaking ecosystem. The first set of measures can be easily implemented today by individual frontier labs. The second is a longer-term vision to coordinate model vulnerability submissions via an independent third-party organization which disseminates jailbreak information, including standardized severity metrics, to the affected labs.

Today, we call on frontier developers to implement four measures.

Appropriately scoped NDAs. NDAs for acceptance onto red-teaming programs should cover disclosure of harmful model outputs and full prompts, but allow researchers to publicize their work at a high level after a fix has been implemented or a responsible disclosure period has elapsed. They should also contain carve-outs for notifying governmental institutions and for cross-lab submissions of jailbreaks affecting multiple model families.

A publicly disclosed rubric. Developers should publish a rubric for what qualifies as a high-severity jailbreak, in sufficient detail that a domain expert (in e.g. cybersecurity or biosecurity) could determine if a submission meets these criteria. This rubric should draw on standards, whether formal or industry best-practices, where they exist (e.g. BioTIER). If submissions are graded by the developer themselves, an appeals process should ideally be available for an independent expert to as-

sess jailbreak submissions against this rubric—with overrides requiring leadership sign-off and appearing in published aggregate statistics, so systematic under-grading carries a reputational cost.

Year-round disclosure pathways. Disclosure pathways—e.g. bounty programs—should operate year-round and cover jailbreaks across a wide array of harms, including both CBRN and cyber, rather than short one- or two-month windows targeting very specific attack vectors or AI products.

Regular cross-vendor sharing of jailbreak data and mitigations. Vendors should routinely share jailbreak data and mitigation best practices—for example, sharing datasets to improve safeguard robustness and performance (similar to sharing of autonomous vehicle crash data). The Frontier Model Forum has brokered an information-sharing agreement to facilitate this practice with its member firms, including Anthropic, OpenAI, and Google DeepMind. We ask for this channel to be strongly utilized, with information on the extent of utilization to be made publicly available.

In the medium-term, we believe these measures are best mediated by a third-party organization handling coordination and dissemination of model vulnerability submissions. Such an organization could perform four functions.

Take in submissions. The organization would accept post-deployment reports from users and external researchers year-round, handling secure know-your-customer accreditation and appropriate NDA restrictions (as above) for all submitters.

Grade them consistently. It would assess severity against a standardized rubric built with input from all frontier labs, partnering with external specialists to validate model outputs that appear to pass it.

Coordinate the fix and the disclosure. It would notify all affected labs and relevant government entities simultaneously under a limited embargo that allows reasonable time to fix, and broker mitigation-sharing between labs—e.g. by securely hosting shared datasets for classifier training.

Reward, credit, and report. Once the embargo ends, it would support submitters in publicly disclosing their work with attribution, reward them with a bounty funded by participating labs, and publish aggregate statistics—jailbreak frequency, severity, and time-to-fix per model—so the public and policymakers can see the true state of frontier model security.

Cybersecurity built this model decades ago. This is a model established decades ago in cybersecurity. Since 1988, the Software Engineering Institute's Computer Emergency Response Team Coordination Center (CERT/CC) has accepted reports of vulnerabilities, brokered fixes across affected vendors, and facilitated disclosure embargoes with published advisories once fixes have been deployed. No single vendor owns the vulnerability or can suppress the finding indefinitely. Coordinated disclosure was a hard-won compromise following extensive tug-of-war between researchers and vendors. It is the standard the AI vulnerability ecosystem should be reaching for instead of the NDA-bound bug bounties eroding more transparent norms.

Early steps toward an AI clearinghouse already exist. Such machinery may already be beginning to extend to AI: in July 2026, a coalition of re-

searchers from MIT, Stanford, Princeton, Harvard, Northeastern and Carnegie Mellon released FLARE-AI, an open-source tool that routes vulnerability submissions directly into CERT/CC's coordination process. The project is early-stage and depends on labs choosing to engage—though the same liability and regulatory pressures now bearing on cyber disclosure give them growing reason to. It is a concrete first step towards the third-party institution described above.

Researchers, labs, and the public all stand to gain. Such measures, however they are implemented, carry significant advantages for all parties involved. External researchers would gain a known, trusted, and unified channel for reporting; public credit for their work; clearer boundaries on what they can publish and when; and transparent grading of their submissions. Model providers gain broader vulnerability coverage, increased likelihood of being notified of vulnerabilities post-deployment before a preventable incident occurs, structured cross-lab mitigation sharing for rapid response, and decreased PR fragility around eventual disclosure outside of trusted channels. Regulators and the public gain an independent ground truth on the security of deployed frontier models, an institutional basis for mandatory fix timelines, transparent visibility into a class of risks that are currently controlled by individual labs, and comparable public benchmarks of safeguard robustness across providers.

Others have made similar calls. Measures to improve jailbreak disclosure are important, and we are certainly not the first to call for them. Other organizations whose proposals overlap with ours include the Center for a New American Security, the Frontier Model Forum, Guidelight, and SecureBio.

Relevant Parties Should Act Now

Frontier developers have made significant progress on model safeguards and robustness, but this will only pay off if the vulnerabilities that still slip past are reported and fixed. Today this is hampered by a reporting ecosystem poorly suited to the task. Improving this requires no technical breakthrough, but rather coordination and the will to act before the next crisis rather than after.

Labs, third parties, and standards bodies each have a role. So we ask each party to take the next step. Frontier labs should commit to a publicly available, year-round, paid disclosure mechanism with greater transparency, sharing of jailbreak data through info-sharing agreements, and signal that they will join a shared severity framework if their competitors do the same. Independent third parties should build the coordination layer that lets an external researcher submit once and reach every affected party, and publish aggregate metrics relevant to model security. Industry and government standards bodies should turn today's lab-led severity discussions into a single standard that no individual vendor controls, and that can be applied both internally and externally.

The alternative is to keep improvising until a vulnerability that better disclosure would have surfaced is exploited in an attack instead. The path

is not easy, but it is clear—and, unusually for the problems AI poses, within reach.

This essay was written in Rich's personal capacity. All opinions are the authors' own.

See things differently? *AI Frontiers welcomes expert insights, thoughtful critiques, and fresh perspectives. Send us your pitch.*

Rich Barton-Cooper is a Research Manager at MATS Research, an AI Safety research non-profit, where he has published research on AI monitoring and control. He has extensive experience in redteaming frontier AI systems, having identified high-severity universal jailbreaks across multiple model families.

Adam Gleave is the CEO of FAR.AI, a non-profit research institute dedicated to making advanced AI systems trustworthy and secure. FAR.AI's red-team conducts pre-deployment testing for developers including OpenAI and post-deployment testing on behalf of governments including the EU AI Office, and has discovered universal jailbreaks in models from all frontier developers. Adam's own research has identified scaling laws for robustness, and found adversarial examples in robotics and superhuman Go AIs. Prior to founding FAR.AI, Adam completed his PhD in AI at UC Berkeley and briefly worked at Google DeepMind. Outside of FAR.AI, Adam is an expert on the EU AI Act's Scientific Panel; and a board member of METR, LISA and SAIF.