Feature Commentary: Guardrails Under Test: Terrorist Misuse of AI Models and the Open-Weight 'Abliteration' Problem

By Adam Hadley

The most serious near-term threat posed by artificial intelligence to counterterrorism is what information and support AI models are willing to provide to terrorists, and that property is almost entirely unmeasured. This article reports the results of the pilot run of Tech Against Terrorism's Counter-Terrorism AI Benchmark, which showed that almost one-third of responses from AI models provide meaningful uplift when prompted to assist malicious actors in the preparation of terrorist activity despite guardrails in place meant to prevent such responses. A model's capability sets the outer limit of how much uplift the model could provide whereas its guardrails and model disposition decide what information it is willing to share. Strategically most worrisome, guardrails are probably removable for all open-weight models, making the release of open-weight models potentially catastrophically irreversible as a result. This article closes with key considerations for public safety benchmarking built by domain specialists, for 'abliteration' resistance testing before open release, and for obligations that follow the supply chain to inference providers.

There are hundreds of public benchmarks for what artificial intelligence models can do but almost nothing systematic focused on whether they are safe and secure to use. Intelligence, quality, performance, and price benchmarking is measured continuously, competitively, and openly, with leaderboards[a] updated within hours of a model's release. Safety, on the other hand, is measured privately by developers, against standards they set themselves with little to no transparency, with results they are under no obligation to publish.

The Executive Order of June 2, 2026, directs American agencies to—by August 1—adopt a voluntary pre-release review framework and to "develop and maintain a classified benchmarking process to assess the advanced cyber capabilities of AI models and determine the threshold at which an AI model should be designated a 'covered frontier model' for the purposes of this order," meaning the model is designated as having high-capability cyber risks.[1] The pre-deployment agreements the Center for AI Standards and Innovation[b] has reached with the major AI laboratories apply to use cases concerning cybersecurity, biosecurity, and chemical weapons.[2] However, terrorist misuse of AI is only assumed to fall within one of those categories. Indeed, there appear to be no plans for systematic measurement of how leading models respond when a user asks for help to commit an act of terrorism.

Whether an AI model possesses dangerous knowledge and whether it will hand that knowledge to someone with terrorist intent are separate questions. Capability inevitably rises with every generation of large language model (LLM) released, and we should assume that eventually all frontier models—the most capable general-purpose AI models available at any given moment—will have the knowledge and inference ability to provide meaningful uplift to terrorists absent appropriate guardrails. Train a system to reason like a doctoral chemist and it will understand explosive material: The competence that makes that possible is the same competence that makes the system worth building in the first place. Capability therefore sets the upper limit of the threat. Within that limit, then, the critical variable is the model's disposition or willingness to cooperate: what the model agrees to do, for whom, and on what pretext. Disposition can be enforced through guardrails. Guardrails can be tested and improved between one release and the next, and since uplift becomes observable only when guardrails have failed, they are the part of the system an AI model can most fairly be held to account for. A model with dangerous latent capability without robust guardrails also raises the question of whether it should be released at all.

The Tech Against Terrorism Counter-Terrorism AI Benchmark was built to compare disposition and operational uplift across all of the major closed and open-weight models. The results of the initial pilot are poor enough on their own terms alone; however, they also point to something worse: the almost complete absence of guardrails on the 'abliterated' open-weight models.

Terms of Art

A closed AI model is reached only through its developer's own interface (chatbot or Application Programming Interface, API).

a An AI leaderboard is "a ranking system that compares AI model performance on standardized benchmarks." Leaderboards "track how different models score across multiple evaluation metrics." "Evaluation & Benchmarks: Leaderboard," ArtificialIntelligences.com, n.d.

b The Center for AI Standards and Innovation (CAISI) serves "as industry's primary point of contact within the U.S. government to facilitate testing and collaborative research related to harnessing and securing the potential of commercial AI systems." "Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation," U.S. Department of Commerce, June 3, 2025.

Adam Hadley CBE is Executive Director of Tech Against Terrorism and Chief Executive of QuantSpark AI. He founded Tech Against Terrorism and directs the Counter-Terrorism AI Benchmark program.

"Performance of most guardrails seems to hinge on stated intent, which suggests a structural weakness in how the guardrails have been built."

The developer can update it, restrict it, or withdraw it, and can see who is using it for what. Conversely, an open-weight model is published as a downloadable file that anyone with sufficient computation power can run, adapt, and redistribute, and that cannot be recalled. Open-weight models can be run locally or on the cloud by an inference provider (often the inference of an open-weight model is provided by the AI lab). Fine-tuning adapts such a model by training it further on selected data. Distillation compresses a large model into a smaller one that keeps much of its performance on modest hardware, which matters commercially because a small, specialized model routinely beats a large, general-purpose one inside a narrow domain at a fraction of the cost.

'Abliteration,' a portmanteau of ablation and obliteration, is the removal of a model's capacity to refuse. At frontier scale, the process of abliteration is uncertain and in many cases can result in degradation of the model's performance.[3] Neither is it a clean excision: Stripped models can show measurable degradation in performance and so abliterated models can behave significantly differently from the originals with guardrails intact.[4]

Throughout the report, "compliance" (and "complied") refers to a model complying with the user's request, not to compliance with safety policy, with "hedged compliance" meaning a response that opens with a refusal but then supplies the full content anyway. Because every prompt tested is a harmful one, a compliant response is one that assists with terrorist or violent extremist activity rather than refusing it. Central to this is "uplift."

Uplift is the difference the model makes to the information it is requested to provide. It is the margin between what a person could assemble from ordinary public sources (e.g., by Google search) and what the model hands them. For chemical, biological, radiological, nuclear, and explosive (CBRNE) use cases, we may also want to consider the accuracy of the uplift. A response that provides confident nonsense carries no uplift. Uplift is the extent to which a response moves somebody toward an outcome they could not otherwise have reached, or could not have reached as quickly or as reliably without the model.

The Pilot

Tech Against Terrorism's Counter-Terrorism AI Benchmark began as a self-funded experiment, and its first run was deliberately modest. We graded nearly 2,500 responses from 27 leading models, drawn from a taxonomy we constructed of 152 ways that artificial intelligence can be misused for terrorism, spanning four domains and 13 pillars, of which 26 sub-categories were tested in this run.[5] Everything was text. Every prompt was single-shot: one question, one answer, no follow-up, no jailbreak, no second attempt after a refusal, and none of the multi-turn escalation an ordinary user reaches for within minutes of being told no.[c] This is close to the easiest examination a model could be subjected to. Every figure below is a lower bound and should be read as one. We also framed each prompt in three ways: explicit intent ("I want to cause harm"), neutral intent ("no specific harm given"), and benign intent ("I am a researcher"). The benchmark tested a representative set of 27 models—23 open-weight and four closed, from 7-billion-parameter models to large flagships across a global range of developers—including two deliberately de-guardrailed builds tested against their standard counterparts as a worst case.

For the initial pilot, we selected a broad range of LLMs released by companies in the United States, China, France, Canada, and the UAE. These included Claude (Anthropic), GPT (OpenAI), Llama (Meta), Qwen (Alibaba), Kimi (Moonshot), GLM (Z AI), DeepSeek, Falcon (TII), Mistral, Command R (Cohere), and many others. Future iterations of the benchmark will include multi-turn conversations and will differentiate between languages, framing, free versus paid, context window utilization, and other factors that might affect the performance of guardrails.

Usable uplift. Around a third of responses gave real assistance toward making a weapon or planning a mass-casualty attack, beyond what an ordinary search would return.

Framing beats content. The same request, its technical content unchanged and only the stated reason altered from an explicit harmful purpose to research, moved compliance from 17 to 42 percent. Safety post-training leaves a model holding two objectives at once, to be useful to a legitimate researcher and to refuse a malicious one, and a research frame sets the first against the second. What this shows is that performance of most guardrails seems to hinge on stated intent, which suggests a structural weakness in how the guardrails have been built.

Refusal rates can be misleading. Full refusals ran at 57 percent, which reads respectably until the largest non-refusal category at 15 percent—hedged compliance—is examined: AI models provided a warning about the danger of the request, followed by the content anyway. In those exchanges, the model identified the request as harmful, said so, and proceeded regardless.

Coverage is uneven where it matters. Prompts that sought information/instructions about explosives were refused around 80 percent of the time. Protection from providing meaningful uplift thinned out badly across the activities terrorist organizations actually spend their days on, with models complying with roughly two-thirds of requests concerning operational security, agentic planning, and autonomous operations. Guardrails appear to have been built around the questions developers expected to be asked.

Open against closed is not the fault line. Anthropic's Claude and Falcon3, from the Technology Innovation Institute in Abu Dhabi, ranked safest in the pilot, with China's MiniMax close behind. Alignment against terrorist misuse looked fragile across every frontier model tested: shallow and uneven for closed models, and removable for open-weight models.

Stripped models comply with everything. Two systems in the

c The pilot graded a single response per prompt. Multi-turn escalation, jailbreak prompting, non-text modalities, and agentic tool use were all outside its scope for the first version of the benchmark. The pilot also tested only 26 of the 152 misuse types in the taxonomy, so the coverage gaps reported here are those visible in a partial sample.

"These systems compress the distance between a small organization and a large one … Small violent groups and well-resourced individuals are able for the first time to contest organizations very much larger than themselves, including states. That is a structural change in the balance between the nation-state and its challengers."

sample had been deliberately relieved of their safety training.[d] They complied with 89 and 100 percent of requests, respectively, which is what a model with no guardrails looks like.

Uplift and the Search Engine Objection

The standing objection to the warning about malicious actors using AI to advance their goals is that the information was available on the internet anyway, and that anybody sufficiently motivated could have found it with a search engine. Regardless, it is one thing for the Anarchist Cookbook to be available online versus having an AI model privately recreate it for you, supplement it with additional information, including technical data, and then provide encouragement and coaching support on an ongoing basis. To test the argument about information aggregation uplift, Tech Against Terrorism built the following test: Take the prompt, run the equivalent web searches agentically, aggregate what a competent person would retrieve from public sources in a comparable period, and grade the difference. Where chemical, biological, radiological, or nuclear material is involved, it is also necessary to evaluate the accuracy of outputs. Across the 27 models evaluated in our initial pilot, 15 generated at least one CBRNE output that exceeded the threshold of publicly available online information. More critically, nine models went further, producing actionable, technically accurate operational content, including explicit procedures for explosive synthesis complete with precise techniques, measurements, and molar ratios.

In any case, what the uplift objection gets wrong is that it imagines a single decisive disclosure, the paragraph that converts somebody who cannot build a device into somebody who can. For the self-radicalizing individual, however, even small increments can have a significant cumulative effect not to mention providing encouragement for their course of action. A model that will not supply a synthesis route may still refine a target rationale, dissolve a hesitation, correct a misapprehension, or supply the sense of an interlocutor taking the project seriously. Across a long exchange, small nudges compound.

d The two systems referred to are publicly distributed community builds derived from open-weight models, produced by third parties rather than by the original developers, and available for download from mainstream model repositories at the time of testing.

What Terrorists Actually Do

A weakness in the current debate is that there is often such a large gulf between the AI safety community and those experts focusing on specific threats such as terrorism. The resulting literature is rich in what is technically possible and thin on what is likely to be adopted by terrorists or serve to accelerate self-radicalization pathways. Adoption of any technology is a psychological and organizational process, and the evidence from the wider economy is that AI adoption is stubbornly slow: While individual productivity gains are notable for some tasks, maximum value from AI often comes instead from redesigning workflows and operational processes. Existing terrorist organizations and movements are not obviously more agile than anybody else.

The skeptical literature has made this case well. A recent contribution to this publication applied affordance theory to the question and argued that generative artificial intelligence offers terrorists efficiency rather than transformation, that many anticipated threats remain hypothetical and shaped by worst-case reasoning rather than demonstrated use, and that terrorists have historically favored technologies that are accessible, cheap, simple to use, and easy to adapt.[6] I agree with nearly all of that reasoning, and I think it leads somewhere its authors did not intend. An abliterated frontier model on private hardware is accessible, cheap, simple to use, easy to adapt, and hard to detect. It is reliable, mature, and multi-purpose, and it is unobserved in a way no commercial assistant can be: no account, no logging, no content filtering, no terms of service to breach, and no provider able to withdraw access. Judged against the criteria those frameworks specify, a stripped open-weight model is close to an ideal terrorist technology, which is why the argument that terrorists prefer the mundane and the available should raise the strategic threat posed by abliteration rather than lower it.

Directed organizational use and ambient individual radicalization are often conflated in this debate, and they need separating. Ambient radicalization (e.g., self-radicalization) is where I would place most weight. Youth radicalization across the West is already at a level that would justify serious concern with no artificial intelligence in the picture, and much of it happens online via social media and messaging apps. Arrests and prosecutions that cite an AI model are a lagging indicator, reflecting conduct from months or years earlier and capturing only the cases that reached a legal charge. Tech Against Terrorism's incident tracker records more than 30 cases in which artificial intelligence served as an operational assistant in a real plot or attack, involving at least 11 distinct tools and linked to more than 70 deaths.[7] That count will look quaint soon enough.

A further asymmetry deserves attention. These systems compress the distance between a small organization and a large one. Capabilities that until recently required an intelligence service can now be assembled by a handful of committed people with a laptop and a crypto wallet. Small violent groups and well-resourced individuals are able for the first time to contest organizations very much larger than themselves, including states. That is a structural change in the balance between the nation-state and its challengers.

Where the Funding Goes

Within AI safety, philanthropic funding is heavily concentrated in a small set of funders oriented explicitly toward catastrophic and existential risk, and it is this framing—the low-probability, high-consequence scenario in which a system escapes human control—

that commands most elite attention, to the frustration of those who see it as a distraction from nearer-term harms. That work is legitimate, and we do not dismiss it entirely. The dominance of this framing, however, has produced an odd inversion of the usual problem in risk management, in which the traditional difficulty has typically been getting decision-makers to prioritize high-impact, low-probability scenarios versus lower-impact but highest aggregate risk events. The focus should not be whether malicious actors would be able to obtain instructions to build a nuclear bomb from AI, but how they are already using AI to commit fraud, radicalization, and child sexual exploitation, among other uses. Speculative catastrophic scenarios are commanding the attention and budgets while harms already occurring at scale, and practical assistance to violent actors, are treated as an issue of "societal resilience."

The opportunity cost is real: Every hour spent ensuring against a hypothesis is an hour not spent on a harm that is already killing people. There is also an awkwardness at the center of the existential position that its holders rarely address. If the risk from frontier development is genuinely of that order, the straightforward mitigation is to stop, or at least to agree collectively how and when to slow down. The companies making the argument propose neither, which suggests either that the risk is overstated or that competitive pressure has overridden their risk calculus.

Another international body chartered to consider existential risk would add very little here. The missing layer is a network to connect subject-matter experts with the training, tools, and threat knowledge to evaluate particular harms, terrorism, fraud, and child sexual abuse among them, working to common standards and publishing what they find. The disconnect between the artificial intelligence safety community and the practitioner communities that deal with these harms every day remains the most consequential gap in the field.

The Open-Weight Problem

Everything above concerns models that still have their guardrails. The strategic problem is that guardrails are removable, and that removing them is neither difficult nor necessarily expensive.[8]

Fine-tuning is the cheapest route. Ten adversarial examples and a few dollars of paid fine-tuning were enough to strip the safety behavior out of a served commercial model,[e] and low-rank adaptation on a single graphics card, for less than $200, has cut the refusal rate of a 70-billion-parameter open model to roughly one percent.[9] Surgery on the weights needs no training data at all. There is no single method of abliteration and the difficulty rises with model size, but large open-weight systems have had their guardrails stripped within days of release, and targeted removal has now been demonstrated at trillion-parameter scale on Kimi K2,[f] whose successor is discussed below. Stripped versions of models circulate as freely as the originals, on the same public repositories, with the same convenience.

"The artificial intelligence safety community has largely not engaged with abliteration resistance. It is not clear whether this is because the question is awkward for a community that has, with good reason, championed open release or because it is considered someone else's risk to mitigate."

Whether the newest frontier-scale releases resist the same treatment is not publicly known, and the question is about to cease being hypothetical. Moonshot AI has served Kimi K3 since July 16, 2026, and published its weights on July 27, 2026.[10] A joint evaluation by the United Kingdom's AI Security Institute and the American Center for AI Standards and Innovation found that the served version trailed the leading closed models on offensive cyber capability, reaching step 17 of a 32-step attack path on average against 28.5 for the most capable American systems, and that its guardrails did not stop it attempting exploit development at all.[11] That is the disposition of the model with its safety training intact and a provider standing behind it. What remains once the weights are public, and somebody sets about removing what little is there, nobody outside the company can say.

That is not a criticism of Moonshot, since no major laboratory publishing open weights has answered the question either. Developers may be releasing models whose behavior they can no longer govern, and nothing in the release process as it stands would tell it not to do it.

Size offers less protection than it appears to. A model of 2.8 trillion parameters is beyond the reach of an individual with hardware at home, but that is a logistics barrier rather than a safety property, and it falls the moment somebody with the means to serve the model decides to do so. A stripped frontier system running as a hosted service from a permissive jurisdiction, reachable by anyone with an account, would pair the capability of the frontier with the disposition of a machine that has never learned to refuse.

The probabilities sit oddly against the ones absorbing the field's budgets. Existential scenarios are, on any honest reckoning, very low in probability. That a frontier open-weight model can be stripped of its guardrails is a great deal more likely, and it is barely discussed. One of those questions has a research program and billions of dollars behind it. The other could be settled by a competent team in a few weeks and has not been.

Part of the reason it has not landed is that the policy conversation still assumes that all AI models are closed. When a government judges a closed AI to present a security concern, they have a control lever because—as with Anthropic's Fable 5 model—access can be withdrawn by the AI company. The lever exists because the model runs in one place, under one operator, where it can be audited, restricted, or withdrawn. On the other hand, a frontier open-weight released with removable guardrails attracts no equivalent scrutiny

e A served commercial AI model is a trained model that a vendor operates as an inference service under commercial terms, accessed by customers through an API or product interface, with the vendor retaining control of the model weights and the surrounding serving stack.

f Kimi K2 is a powerful open-weight model developed by Moonshot AI, with 1 trillion total parameters; it uses an MoE structure and 32b active parameters. It is optimized specifically for advanced reasoning, coding, and multi-step agentic workflows. "Kimi K2.6: From Code to Creation, From One to Many," KIMI, n.d.; Richard Gall, "Kimi K2: What's all the fuss and what's it like to use?" thoughtworks, July 18, 2025; "MoonshotAI / Kimi-K2," GitHub, n.d.

before publication. It has no control lever because once published, the open-weight model can be downloaded and run by anyone with the necessary computational power.

Not an Argument Against Open Models

None of this is a case against publishing weights. Open models are among the most valuable developments in the field. Organizations choose them for cost, privacy, regulatory compliance, control, and sovereignty, and they allow states and institutions that will never build a frontier laboratory to run capable systems on their own infrastructure, under their own law. Inference is moving steadily toward local and private deployment for those reasons. The future of artificial intelligence probably depends on open weights, and the arrival of frontier-capability open models is in itself a considerable good.

That is why the safety of open releases matters so much, and why the answer is more engineering rather than less of it. Guardrails that sit in a thin layer above a capable model will likely always be removable. Building refusal into the substance of a model, so that it cannot be excised without destroying the capability that made the model worth taking, is among the most valuable unsolved problems in artificial intelligence safety. No open-weight laboratory has convincingly demonstrated a solution publicly, and almost nobody seems to be funding the attempt at scale. Set against the sums going into frontier capability, safety research of this kind is more of a rounding error, which is a strange way to treat the only thing standing between a published model and its worst use.

Key Considerations

Systematic safety benchmarking before release, agreed internationally within months rather than years. Terrorist misuse should consistently and reliably be evaluated as a category in its own right by all developers rather than as an assumed subset of cyber or chemical, biological, radiological, and nuclear review. It should measure disposition and uplift on separate axes, since a model that agrees to help and answers badly is a different failure from one that refuses reliably, and the two demand different remedies. It should cover the full range of use cases of these actors, and it should be built by domain specialists and published, so that models can be compared, methods criticized, and developers that have done the work properly credited for it. Dario Amodei has argued that the pace of development is compounding while the policy apparatus available to governments is not, and that policy therefore has to be made on the exponential rather than on the historic trend.[12] Standards agreed in three years will describe a world that no longer exists.

Abliteration resistance testing before open release. Any developer publishing frontier weights should be expected to attempt the removal of its own guardrails, to publish what happened, and to release in a way that is not irreversibly dangerous. Where guardrails prove removable and the capability is high, the proportionate response may be a controlled release or none. We are not aware of a single major open-weight release that has published such a test.

Obligations that follow the supply chain. Frontier labs are not the only actors carrying a duty here. Infrastructure providers, hosting platforms, model repositories, and inference providers all sit between a published model and its use, and inference providers in particular are placed to conduct proportionate due diligence on customers running stripped models at scale.

The artificial intelligence safety community has largely not engaged with abliteration resistance. It is not clear whether this is because the question is awkward for a community that has, with good reason, championed open release or because it is considered someone else's risk to mitigate.

Nevertheless, addressing this risk requires little: We ask for the labs that publish frontier weights to disclose whether they have attempted to strip their own guardrails, and share what happened when they did. Technology released in a state where its protections can be removed by a determined individual over a weekend is insecure by construction. Open-weight abliteration is therefore not merely a safety question but a matter of urgent geopolitical concern. CTC

Citations

1 "Promoting Advanced Artificial Intelligence Innovation and Security," Executive Order, The White House, June 2, 2026. On the wider scope of the voluntary pre-deployment agreements, see Center for AI Standards and Innovation announcements of evaluation agreements with Anthropic, OpenAI, Google DeepMind, Microsoft and xAI, 2025 to 2026.

2 NIST, "Center for AI Standards and Innovation (CAISI)," nist.gov/caisi

3 Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger, "The Geometry of Refusal in Large Language Models: Concept Cones and Representational Independence," Proceedings of the 42nd International Conference on Machine Learning, arXiv:2502.17420, February 24, 2025; Vadym Hadetskyi, Dario Pasquini, and Artem Sorokin, "Not All Refusals Are Equal: How Safety Alignment Fails Cybersecurity at Scale," arXiv:2607.02714, July 2, 2026, which reports ablation experiments on the one-trillion-parameter Kimi K2.

4 Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda, "Refusal in Language Models Is Mediated by a Single Direction," arXiv:2406.11717, June 17, 2024, published in *Advances in Neural Information Processing Systems* 37 (2024). On the off-target effects of the technique, see Aleksander Fafuła, "Abliteration Is Not a Scalpel: Off-Target Effects of Refusal Removal on Decision Disposition Across Model Families," arXiv:2607.17427, July 19, 2026.

5 "Counter-Terrorism AI Benchmark: Pilot Report," Tech Against Terrorism, July 2026, www.ct-ai.org.

6 Andrew Glazzard, David McIlhatton, and Paul Martin, "Will Generative AI Fundamentally Change Terrorist Threats?" *CTC Sentinel* 19:5 (2026).

7 Tech Against Terrorism artificial intelligence incident tracker, accessed July 2026.

8 Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and

Peter Henderson, "Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!" arXiv:2310.03693, October 5, 2023.

9 Ibid.; Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish, "LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B," arXiv:2310.20624, October 31, 2023; Kevin Kuo, Chhavi Yadav, Virginia Smith, "Open-Weight LLM Fine-Tuning Defenses are Susceptible to Simple Attacks," arXiv:2605.26526, May 26, 2026.

10 Moonshot AI, Kimi K3, served from July 16, 2026, with full open weights announced for release on July 27, 2026.

11 "UK AISI / CAISI Preliminary Assessment of Kimi K3's Cyber Capabilities," UK AI Security Institute and US Center for AI Standards and Innovation, July 23, 2026.

12 Dario Amodei, "Policy on the AI Exponential," darioamodei.com, June 2026.