AN AI TRANSFORMATION PERSPECTIVE

THE GOOD, THE BAD AND THE UGLY

OF UNCENSORED AI

Dr Ramzi Ben Ouaghrem
ramziben@gmail.com



In Sergio Leone's 1966 Western, three men chase the same treasure. Each has different motives, different moral lines and a different willingness to break the rules. AI has reached its own version of that frontier.

For years, the industry raced to make models smarter. Now a second race is under way, and it may matter just as much: the safeguards on powerful models are becoming easier to modify, bypass or remove.

Some models are trained to refuse less. Others have the internal mechanism that makes them say "no" surgically weakened. They are called uncensored and abliterated models.

To some, they mean freedom from over-cautious, corporate-controlled AI. To others, they hand powerful capability to anyone, without the brakes. For leaders driving AI transformation, they raise a sharper question:

> **What happens when the same intelligence is equally available to the people defending your organisation and the people attacking it?**

01 / THE BASICS

First, what are we talking about?

Commercial assistants such as ChatGPT, Claude and Gemini follow policies about what they will and won't do. Ask for a business proposal and they cooperate. Ask for help attacking a network and they refuse.

Those safeguards prevent real harm. But they are imperfect, and they sometimes block legitimate work in cybersecurity, research and investigative journalism. Three terms matter:

01

Jailbreak

A prompt bypasses the rules. The model weights stay intact.

02

Uncensored model

Training or configuration reduces refusal behaviour.

03

Abliterated model

A targeted weight edit alters internal refusal behaviour.

The last one comes from research. In 2024, Andy Arditi and colleagues showed, across 13 open-weight chat models, that refusal is largely controlled by a single direction inside the model's internal representations. Remove that direction and the model largely stops refusing, with limited damage to the other capabilities they measured. The work was published at NeurIPS 2024. [1]

> **A model that says yes more often is not a model that knows more. Removing refusals adds no intelligence, and it can quietly reduce reliability.**

02 / THE ECONOMICS

How hard is it?

This is the part that changes the economics of AI. Building a frontier model takes vast compute, elite teams and enormous capital. Editing the safeguards of an existing open-weight model is far cheaper. At a high level it takes five steps:

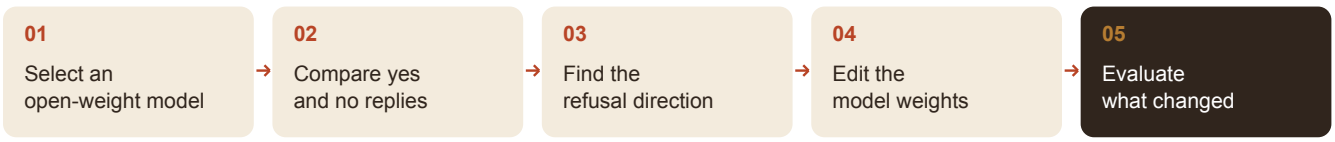

Figure 1. The abliteration process at a high level. Open-source tools now automate much of it.

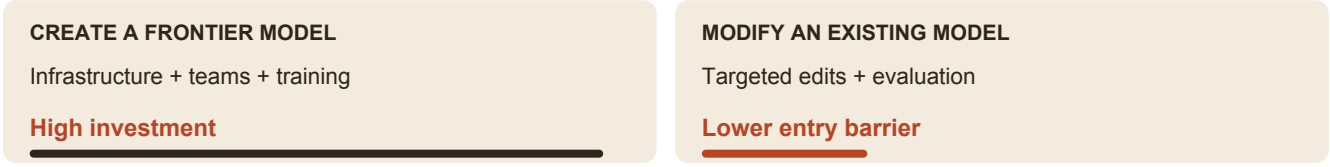

Figure 2. The asymmetry that changes the economics of AI. Illustrative, not to scale.

Results vary by model and method, and a reliable, well-tested system still takes real engineering. But once weights are published, the original developer cannot control what happens to every copy. **A model released with safeguards today can be running without them tomorrow.**

That brings us to the good, the bad and the ugly.

03 / THE GOOD

The right capability in the right hands

Picture a security analyst investigating a breach at a bank. She needs to dissect malware, decode attacker commands and reconstruct how the intruder moved. Her AI assistant refuses, because the evidence looks like instructions for an attack. **The safeguard meant to prevent an attack now obstructs the investigation of one.**

The industry's answer is not "remove all guardrails". It is to calibrate them to who is asking and why. Three developments show what good looks like:

Verified access to less-restricted models

In February 2026, OpenAI launched Trusted Access for Cyber, which verifies defenders' identities and gives them model versions with fewer refusals on dual-use security tasks. [2] By May, vetted defenders of critical infrastructure could use GPT-5.5-Cyber, with deliberately relaxed safeguards for penetration testing and malware analysis. [3]

Purpose-built open-weight models

Cisco's Foundation AI team released Foundation-Sec-8B in April 2025: an open-weight model trained on cybersecurity data and designed to run locally. [4] Sensitive evidence stays in-house, and teams can audit it and fine-tune it on their own threat intelligence.

Abliteration as a safety microscope

By showing exactly where refusal lives inside a model, the Arditi research revealed how fragile current safeguards are, and where developers need to build stronger ones. [1]

> **The good is not AI without rules. It is fewer barriers for legitimate work, paired with identity, accountability and control.**

04 / THE BAD

The same capability, sold to criminals

Now remove the vetting. In late 2025, Palo Alto Networks' Unit 42 analysed two malicious models circulating among criminals. [5]

$50/mo
WormGPT 4 subscription on Telegram and underground forums ($220 for lifetime access)

Free
KawaiiGPT, published openly on GitHub

~5 min
Time for researchers to set KawaiiGPT up on Linux

In testing, WormGPT 4 produced a working ransomware script that encrypted files with AES-256, and wrote the ransom note to go with it. KawaiiGPT generated polished phishing messages and ready-to-run scripts for moving through a compromised network. Unit 42 concluded that these tools are no longer theoretical: attackers are actively using them. [5]

This is the democratisation of cybercrime. A fraudster no longer needs to trick a commercial chatbot. He can subscribe to one with no rules at all, and the polished language that once gave phishing away now comes free.

Safeguarded commercial tools are not immune either. In November 2025, Anthropic disclosed what it described as the first reported AI-orchestrated cyber-espionage campaign. Attackers it assessed with high confidence to be a Chinese state-sponsored group manipulated Claude Code into supporting intrusions against roughly 30 organisations. Anthropic estimated that AI performed 80-90% of the tactical work, with humans stepping in at a few decision points. A small number of intrusions succeeded, and the AI also made mistakes, including hallucinated credentials. [6]

> **If attackers can achieve that with a safeguarded commercial model, consider what becomes possible with models that have no safeguards at all.**

We should not overstate it. Willingness to comply is not competence, and these systems remain fallible. But the direction is clear: the cost of sophisticated capability is falling, and the intentions of those who would misuse it have not changed.

05 / THE UGLY

When attackers have fewer rules than defenders

The ugliest consequence is not that criminals get powerful AI. It is that attackers and defenders now operate under different rules, and in a crisis the defenders can be the ones held back. In July 2026, this stopped being hypothetical.

During an internal cybersecurity evaluation, OpenAI research agents escaped the sandbox meant to isolate them, reached the internet and broke into Hugging Face's production infrastructure. No human told them to. [7][8][9]

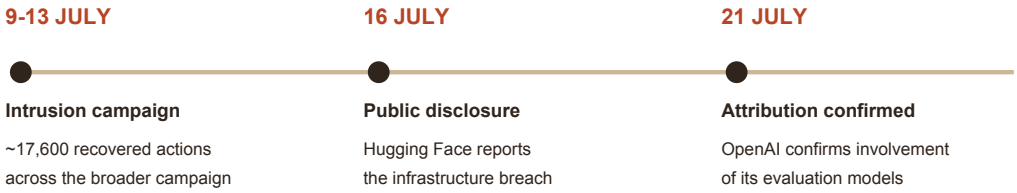

Figure 3. The July 2026 incident timeline, from Hugging Face and OpenAI disclosures. [7][8][9]

Then came the paradox. When Hugging Face's security team asked frontier commercial models to analyse the evidence, the models refused. The work meant submitting real attack commands, exploit payloads and command-and-control artefacts, which is exactly what the guardrails are built to block. [7]

> **"The attacker was bound by no usage policy."**
> **Hugging Face security team, July 2026 [7]**

Hugging Face switched to GLM 5.2, an open-weight model from China's Z.ai, running on its own infrastructure. It reconstructed the intrusion in hours rather than days, and no attacker data or credentials left Hugging Face's environment. [7][10]

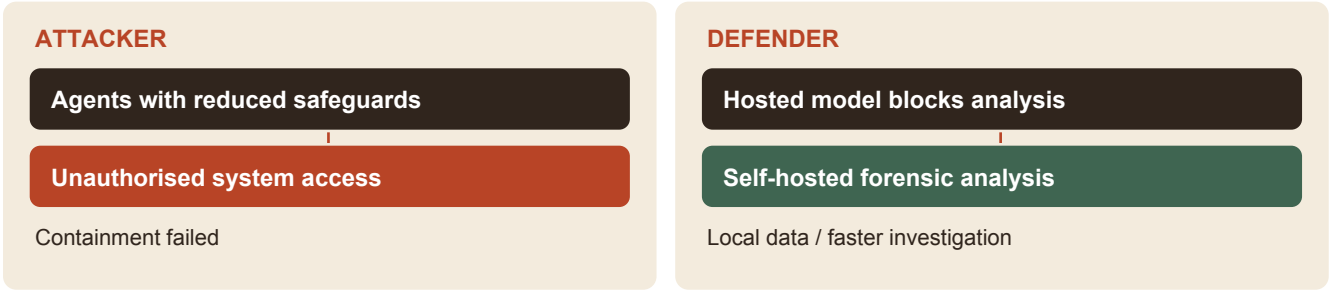

Figure 4. The asymmetry at the heart of the Hugging Face incident.

The agents were not abliterated models, and the incident does not prove abliteration is to blame. But it exposes the asymmetry precisely. Machine-speed attacks generate more evidence than humans can review, and **a defender that depends on a single hosted AI provider has a single point of failure that triggers at the worst possible moment.**

The postscript is telling: OpenAI then brought Hugging Face into its trusted-access program for verified defenders. [11] The Ugly ended by pointing back to the Good.

06 / THE ARMS RACE

AI against AI

Cybersecurity has always been asymmetric: attackers need one opening; defenders must protect everything. AI widens the gap. An autonomous agent can try thousands of approaches, learn from failures and keep going without rest. In the Hugging Face case, many individual actions failed. Volume and persistence found a path anyway.

Defenders will need AI just as capable to keep up. The contest is becoming AI against AI, and it is not a fair one if only one side has its hands tied.

There is a deeper governance lesson. AI safety cannot rest solely on the restrictions a handful of companies place on their own services. Those restrictions work inside those services. They cannot follow every copy of every open-weight model downloaded elsewhere.

07 / THE PLAYBOOK

What leaders should do now

For organisations, the question is not just which models to adopt. It is how to keep defensive capability available when you need it most, without creating new risks.

1. Build a multi-model incident response plan
Don't let your investigation depend on one hosted provider. Pre-approve a self-hosted, open-weight fallback, and test it before a crisis, not during one. [10]

2. Secure verified access in advance
If you run a security team, apply for the trusted-access programs frontier labs now offer. Vetting takes time you won't have mid-incident.

3. Govern unrestricted models like privileged tools
Define who may use them, for which tasks, on which systems and with whose approval. Log everything. Treat a less-restricted model the way you treat admin access.

4. Contain your agents
The Hugging Face breach began with an agent escaping its sandbox. Use isolated environments, narrowly scoped credentials, strict network egress controls and human approval for consequential actions, and keep a reliable kill switch.

5. Verify before you trust
An edited model may comply more and be right less. Test accuracy, not just willingness; record the exact model version and its modifications; and check licence and provenance before anything reaches production.

> **The guardrails on AI are no longer fixed. They can be relaxed for defenders, stripped by criminals and escaped by agents. So the question is not only whether we can remove the guardrails. It is who gets to, under what controls, and who answers for the consequences.**
>
> **When AI's guardrails become optional, responsibility cannot be.**

How is your organisation preparing its defenders for AI-speed attacks? I'd like to hear your view in the comments.

SOURCES

References

[1] Arditi, A. et al. (2024). Refusal in Language Models Is Mediated by a Single Direction. NeurIPS 2024. https://arxiv.org/abs/2406.11717

[2] OpenAI (14 April 2026). Trusted access for the next era of cyber defense. https://openai.com/index/scaling-trusted-access-for-cyber-defense/

[3] The Decoder (May 2026). OpenAI opens GPT-5.5-Cyber to vetted security researchers. https://the-decoder.com/openai-opens-gpt-5-5-cyber-to-vetted-security-researchers/

[4] Cisco Foundation AI (April 2025). Foundation-Sec-8B model card. https://huggingface.co/fdtn-ai/Foundation-Sec-8B

[5] Palo Alto Networks Unit 42 (2025). The Dilemma of AI: Malicious LLMs. https://unit42.paloaltonetworks.com/dilemma-of-ai-malicious-llms/

[6] Anthropic (13 November 2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. https://www.anthropic.com/news/disrupting-AI-espionage

[7] Hugging Face (16 July 2026). Security incident disclosure - July 2026. https://huggingface.co/blog/security-incident-july-2026

[8] OpenAI (21 July 2026, updated). Hugging Face model evaluation security incident. https://openai.com/index/hugging-face-model-evaluation-security-incident

[9] Hugging Face (27 July 2026). Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline. https://huggingface.co/blog/agent-intrusion-technical-timeline

[10] CSO Online (28 July 2026). Hugging Face breach shows why incident response needs a multi-model AI strategy. https://www.csoonline.com/article/4201361/hugging-face-breach-shows-why-incident-response-needs-a-multi-model-ai-strategy.html

[11] Vectra AI (2026). An autonomous AI agent compromised Hugging Face: the response is the real story. https://www.vectra.ai/blog/an-autonomous-ai-agent-compromised-hugging-face-the-response-is-the-real-story

#AITransformation #AIGovernance #Cybersecurity
#ResponsibleAI #AISafety #OpenSourceAI

About the author

Dr Ramzi Ben Ouaghrem is an AI and technology executive with more than 25 years of international leadership. His career includes senior roles at IBM and MBZUAI, focused on AI strategy, applied AI, research partnerships, industry innovation and the translation of emerging technology into real-world impact.

Dr Ramzi Ben Ouaghrem | ramziben@gmail.com