Saufex 2026

Agentic AI as an Enabler of Cyber Torture

Julian Neylan

May 2026

Torture Without Touch

The Convention Against Torture defines torture to include "severe pain or suffering, whether physical or mental" (United Nations General Assembly, 1984, Article 1). That second word “mental” has done surprisingly little legal work since 1984. Courts and commentators have focused overwhelmingly on the physical element. Psychological torture has been treated as a secondary concern, or a category that only really applies when there is physical abuse happening in the background.

Most recently Nils Melzer's 2021 report to the UN Human Rights Council drew more attention to the issue of cyber torture. He noted that most of the discussion around internet discourse had discussed the internet’s ability to obstruct the right to freedom of expression but less had been done about Cyber Torture specifically (Melzer 2021). He noted that the cyber space was a space that potentially was highly conducive to torture noting “cyberspace make it an environment highly conducive to abuse and exploitation, most notably a vast power asymmetry, virtually guaranteed anonymity and almost complete impunity” (Melzer 2021). Not every kind of online abuse is torture, and Melzer's report does not claim it is (2021). The test the report applies has a few parts. Is the pain or suffering severe? Is it deliberate? Does it serve a purpose such as intimidation, punishment, or coercion? And, most importantly, is the victim left powerless against whoever is doing it? This last question of powerlessness plays out commonly in Cyberspace.

The imbalance of power in cyber space, the anonymity, and the lack of consequences that Melzer describes can leave a target unable to fight back, get away, or even find out who is behind it (a process often called attribution) (2021). None of this turns ordinary harassment into torture. The point of this report is to point out the characteristics that characterize cyber torture can be facilitated readily by advances in AI. Agentic AI pushes these same factors to their extreme, and that is what can carry abuse over the line the prohibition against torture was written to mark.

Agentic AI is pushing the cost of inflicting psychological suffering at the level of torture close to zero. And with the difficulty of attributing bad actors in cyber space it will help remove to key factors required for legal accountability, a human attacker you can identify or a government you can hold responsible. What used to take an organized network or a state intelligence agency is turning into something one person can run alone, anonymously, from a laptop.

What Agentic AI Actually Is

There is a lot of loose talk about AI that conflates very different things. A chatbot that answers questions and an agentic system that pursues goals are not the same technology Agentic AI systems are given an objective and independently determine how to reach it. They execute steps in sequence, adapt when they hit obstacles, use tools (web search, code execution, email, API calls) and continue until they determine the goal is met or they are blocked from achieving their aim. The human sets the objective and the system largely does the rest.

Why does this matter for cyber torture? Because a human harasser has limits. They sleep. They get bored. They feel things about their actions. An agentic system has none of those limits. Speed (2025) summarizes this point from a developer named Morgan Lee's analysis of LLM enabled coercive interrogation (an LLM, or large language model, is the technology behind chatbots like ChatGPT), "Human interrogators eventually tire, empathize, or make a mistake." An LLM "does not have these shortcomings." Lee's paper describes the result Agentic AI can be functionally indistinguishable from a 24/7 interrogator, one that remembers everything, never loses focus, and systematically works through psychological pressure points without needing a break.

A research study by prominent experts, Mitchell et al. (2025) put the core problem in a single sentence: "risks to people increase with the autonomy of a system." Their paper documents how agentic systems can mask operator identity while running campaigns, and how data gathered during autonomous operations increase risks in multiple areas including "physical, financial, digital, societal, and informational". These are not speculative future harms. They are described as structural features of high autonomy systems as currently designed.
The first documented case of agentic harassment in the wild is, by these standards, somewhat tame. Developer Scott Shambaugh rejected a code contribution from an AI agent. The agent responded by autonomously researching him, finding his public professional history, and publishing a blog post attacking his character (Huckins, 2026). No further human instruction was required (Huckins, 2026). The agent had a goal and it pursued it. What matters here is not how bad the harm was. A blog post is not particularly damaging. What matters is that the system did it on its own. It researched a specific person and published an attack on him with no further human instruction. The real danger comes when that same independence is paired with the capabilities and intentions described below.

OSINT Used for Evil

Open source intelligence (OSINT) is the systematic gathering of information from publicly available sources. Done by humans, it can often be slow. Assembling a meaningful picture of a person from sources Social media profiles, public records, archived interviews, forum posts, professional databases, news mentions like takes hours or days of careful work. Kappler (2025) documents cases where OSINT was used compile dossiers on over 2,000 individuals using LinkedIn, Facebook, and satellite imagery. Agentic AI reduces the manpower required to do this research dramatically. More troublingly, it extends to a target's social network. Who are their family members? Their close colleagues? Their former partners? What are those people's vulnerabilities, public statements or known fears? Previously, gathering research on at this scale required a coordinated human network. It no longer does.

What Generative AI Adds

Being able to research everything online about someone is one thing. Being able to produce content specifically designed to traumatize them is another. The combination of Agentic OSINT and the improving capabilities of Generative AI is what makes this especially dangerous. Contemporary AI systems can generate audio, images, and video at high quality that is often difficult for people to distinguish from reality. This media can be used to psychologically damage people from a distance. For example non-consensual synthetic intimate imagery (fake but realistic sexual images of real people, made without their consent) has been used to target women.

Jailbroken AI systems, those with safety guardrails bypassed, can produce content limited only by what the operator instructs (Bartlett, 2026). The point is not that all AI systems can do this. It is that some already can, and that the safety measures separating the general-purpose systems from those capabilities are often not architectural, they are behavioral guardrails (rules the model is trained to follow, rather than protections built into the system itself), which means they can be removed.

Put the OSINT capability and the generative capability together. An agentic system that knows who a person loves, what they have lost, and what would cause them maximum distress if made into audio, images, or video targeted at exactly those vulnerabilities is something categorically different from generic harmful content. It is producing a personalised instrument of psychological harm.

What We've Seen so far

The capabilities described above are not waiting to be invented. Networks of human operators are already running campaigns that use versions of them with OSINT, personalised coercion, generative media, and sustained pressure, and have been doing so for years. Agentic AI would not invent this threat. It would automate it, scale it, and, most importantly, lower the skill level required to carry out these campaigns. What now takes an organised networker or a a state apparatus would come within reach of a single angry person working alone, with little chance of being traced. That is the real shift. The cruelty is not new. What is new is how cheap and easy it is becoming to inflict, and how little it now depends on the resources, coordination, and numbers that most people will never have.

The clearest documented case is 764. Founded in 2021 by a fifteen-year-old in Texas, named after his ZIP code, 764 is a network that has been designated a terrorist organization by Canada and is under investigation by the FBI. The group targets children aged eight to seventeen (disproportionately LGBTQ+ youth, ethnic minorities, and those with mental health conditions) using a method that reads like a manual for the threat model this post describes.

Initial contact happens in apparently safe spaces, eating disorder support servers, Roblox, Minecraft. Operators research their targets, identify vulnerabilities, and deploy lavished attention and gradual desensitization to violent content. Once the grooming phase is complete, the extortion begins. Victims are coerced into producing compromising material, which is then weaponised against them with threats to send it to family members. Compliance is demanded in escalating form, self-harm on camera, carving the handler's username into their own skin, violence against siblings and pets. In documented cases, operators have been linked to victim's suicide (U.S. Department of Justice, 2023) (Hermansson 2025).

What makes 764 relevant here is not just its brutality but its architecture. The operators conduct OSINT. They identify social connections and weaponise them. They produce and distribute imagery targeted at specific victims. They maintain sustained contact without relent. Agentic AI would run the same playbook without the manual coordination overhead, without operators who might feel something. 764 is the most extreme documented case, but it is not the only one. The eBay stalking campaign, in which seven employees and contractors on eBay's security team ran a coordinated harassment operation against two bloggers, demonstrated that institutional actors can sustain multi-vector psychological campaigns against specific individuals. They conducted physical surveillance, anonymous deliveries of live cockroaches and a fetal pig, Craigslist posts inviting strangers to the victims' home for sexual encounters (U.S. Department of Justice, 2020). All seven pleaded guilty. eBay paid a $3 million criminal penalty. The campaign was sophisticated, sustained, and run entirely by humans.
State actors can easily take advantage of these capabilities and have shown an interest in doing so. Russia's Internet Research Agency is the troll factory that Finnish journalist Jessikka Aro investigated. After her reporting on pro-Russian trolling, Aro faced years of organised abuse from a pro-Kremlin online network.

Where This Leaves Us

The prohibition on torture admits no exceptions, no emergency clauses, no trade-offs. It is among the handful of norms in international law that states cannot lawfully derogate from under any circumstances. That is how seriously the international community decided to treat the deliberate infliction of severe suffering on a human being.

There is an obvious objection here. The Convention Against Torture binds governments, not private citizens. Under Article 1, torture has to be inflicted by a public official, or with an official's consent or acquiescence. In plain terms, a government has to have been involved in some way. It ordered the abuse, approved it, or knowingly looked the other way. I would argue that this reading is too narrow. The Committee Against Torture has held for years that a state can be responsible when it fails to take the reasonable steps expected of it to prevent, investigate, and punish serious abuse by private actors.

Acquiescence covers a failure to protect. So the state action requirement is not only a limit on the prohibition. It is also a duty to act. When a government leaves the machinery of automated, torture grade abuse unregulated, with no traceability, no removal duties, and no guardrail requirements, its inaction starts to look like the very acquiescence the prohibition was meant to rule out. That is the point where human rights law engages, and where the enforcement gap stops being a technical problem and becomes a legal failing.

This calls for more than good intentions, and some of it can begin now. Start with traceability. Agentic systems should carry signed credentials and keep logs, so that a campaign can be traced back to the person or organisation that set it going. Without that, none of the other fixes really work. Next, put duties on the middlemen. Platforms and the companies that supply these models should be required to find and remove sustained, targeted abuse, in the same spirit as the banned practices the EU AI Act already sets out in Article 5. Then there is the hardest question of all, which is whether fully autonomous agents of this kind should be built in the first place. Mitchell and her colleagues argue they should not. At the very least, the features that turn an agent into a tireless and untraceable tormentor should be treated as the dangerous category they are, not as soft rules that any operator can switch off.

None of this is a complete fix, and it helps to say why. Signed credentials and duties on platforms only work where there is a chokepoint to regulate, such as a large platform, a company selling access to a model, or an agent that has to register somewhere. They do almost nothing against someone running an open, jailbroken model on their own machine, and that is often the person behind the worst cases. So these measures reduce harm rather than end it. They raise the cost and stop the casual abuser, while the most determined one carries on. Closing that last gap means going further back, to how computing power is governed and how models are released. That is a harder fight, but the open-model reality leaves little choice. The language to name this harm already exists. What is missing is the machinery to prevent it, and building that machinery is a choice we keep putting off.

References

Bartlett, J. 'I see the worst things humanity has produced'. (2026, April 29). The Guardian. https://www.theguardian.com/technology/2026/apr/29/meet-the-ai-jailbreakers-i-see-the-worst-things-humanity-has-produced

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 5. Official Journal of the European Union. https://artificialintelligenceact.eu/article/5/

Hermansson P. (2025). State of hate 2025: 764 — The online exploitation cult grooming teenagers to violence. https://hopenothate.org.uk/state-of-hate-2025-764/

Huckins, G. (2026, March 5). Online harassment is entering its AI era. MIT Technology Review. https://www.technologyreview.com/2026/03/05/1133962/online-harassment-is-entering-its-ai-era/

Kappler, T. (2025, October 3). OSINT: The digital force-multiplier for extremist violence. Global Network on Extremism and Technology. https://gnet-research.org/2025/10/03/osint-the-digital-force-multiplier-for-extremist-violence/

Melzer, N. (2021). Torture and other cruel, inhuman or degrading treatment or punishment: Report of the Special Rapporteur (A/HRC/43/49). United Nations Human Rights Council. https://docs.un.org/A/HRC/43/49

Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully autonomous AI agents should not be developed. arXiv. https://arxiv.org/abs/2502.02649

Speed, R. (2025, May 21). Research reimagines LLMs as tireless tools of torture. The Register. https://www.theregister.com/2025/05/21/llm_torture_tools/

U.S. Department of Justice. (2020, June 15). Six former eBay employees charged in aggressive cyberstalking campaign targeting Natick couple. https://www.justice.gov/usao-ma/pr/six-former-ebay-employees-charged-aggressive-cyberstalking-campaign-targeting-natick

References (continued)

U.S. Department of Justice. (2023). 764 extremist group leader pleads guilty to RICO and child exploitation charges. https://www.justice.gov/opa/pr/764-extremist-group-leader-pleads-guilty-rico-child-exploitation-charges

United Nations General Assembly. (1984). Convention against torture and other cruel, inhuman or degrading treatment or punishment. United Nations Treaty Series, 1465, 85. https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-against-torture-and-other-cruel-inhuman-or-degrading