

# 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 a secondary concern, or a category that only really applies when there is physical abuse happening in the background.

Most recently Nils Melzer's 2020 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 2020). 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 2020).

That was 2020. Melzer was describing a world of human-operated harassment campaigns and surveillance tools. Since then we've seen large advancements in technology that can further capitalize on the anonymity and power asymmetry capable on the internet. Specifically, through Agentic AI we see that the potential for harm is greatly expanded.

## 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, "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. It pursued it. That is the floor, not the ceiling.

## 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 to compile dossiers on over 2,000 individuals using LinkedIn, Facebook, and satellite imagery work that previously required an intelligence apparatus.

Agentic AI compresses that 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 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 traumatise them is another. The combination is what makes the combination of Agentic OSINT and the improving capabilities of Generative AI 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 has been used to target women. Citron (2022) recorded over 50,000 deepfake sex videos in circulation at time of publication, nearly all targeting 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, 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.

# This Is Not Hypothetical: 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 create this threat from scratch. It would automate and scale what already exists.

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 desensitisation 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 senior employees ran a coordinated harassment operation against two journalists, 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, which ran coordinated harassment operations targeting journalists and activists, including Finnish journalist Jessikka Aro, who was subjected to years of organised abuse after reporting on pro-Russian trolling.

## 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.

Agentic AI creates a new delivery mechanism for that suffering. It is not hypothetical. The documented cases are early and mild by comparison with what the technology can already do. What we have is autonomous systems that can research a target and their social network without fatigue, generate personalised content designed to traumatise them specifically, deliver that content across every digital channel simultaneously, never stop, and be operated by someone who may be functionally untraceable.

The human rights framework exists to cover exactly this. The enforcement infrastructure does not yet exist to apply it. Attribution of who is held responsible in these cases is made more difficult when agentic AI is utilized. We need to build the infrastructure with traceability requirements for agentic systems. Platform removal obligations tied to harm need to be taken seriously. We need to build better guardrails for both Agentic and generative AI systems.

## References

- Bartlet 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>
- Citron, D. K. (2022). *The fight for privacy: Protecting dignity, identity, and love in the digital age*. W. W. Norton.
- 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. (2020). *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/](https://www.theregister.com/2025/05/21/llm_torture_tools/)

## References (continued)

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>

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>