Session II: Wellness and Wellbeing in the age of AI

Revisiting the Dark Web: GBV, CSAM and Agentic AI

Dr. Jumoke Giwa

Revisiting the Dark Web: GBV, CSAM and Agentic AI

Artificial Intelligence is now a double-edged sword capable of automating extreme digital exploitation while simultaneously offering sophisticated tools for victim detection.

"Agentic AI" acts as a force multiplier. In the deep and dark web, this autonomy is weaponized.

Dark Web

Hidden internet requiring Tor.

Not inherently criminal; used for privacy, e.g. protecting journalists; but also attracts criminal networks.

GBV

Gender-Based Violence.

Harmful acts directed at individuals based on gender.

Amplified by digital tools.

The spectrum now includes doxxing, non-consensual sharing of intimate images, and AI-generated fake sexual imagery (deepfakes). The harm is psychological, reputational, and sometimes permanent.

CSAM

Child Sexual Abuse Materials.

Content depicting sexual exploitation of children. Distributed via encrypted platforms.

Possession and distribution are severe felonies worldwide. This is a top priority for global law enforcement and must be distinguished from general “adult content" debates.

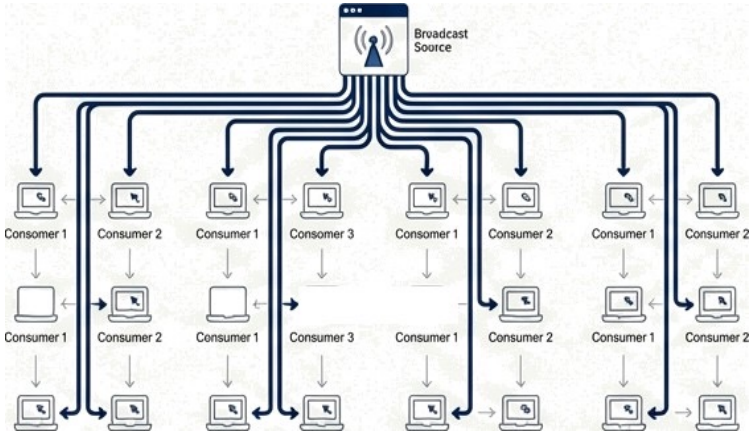
Agentic AI

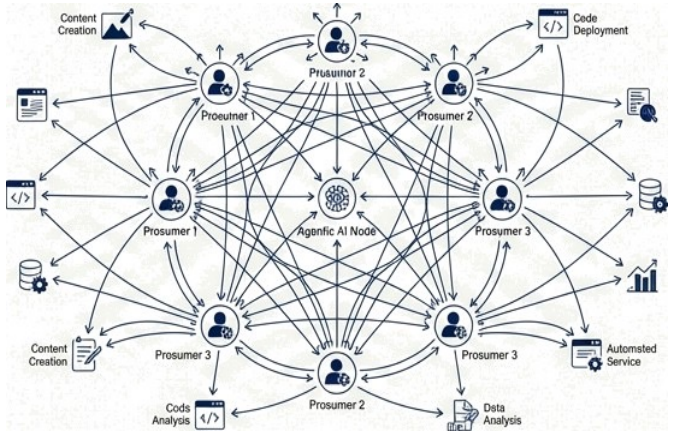
Autonomous systems that use AI agents – powered by large language models (LLMs) and specialized tools – to plan, reason, and act independently to achieve complex, multi-step goals with limited human supervision.

Unlike generative AI that creates content upon request, agentic AI focuses on decision-making, interacting with applications, and adapting to dynamic environments.

It's acting on behalf of the owner/creator.

Shift from passive consumers to active prosumers and producers.





The Convergence:
where technology and harm intersect

-

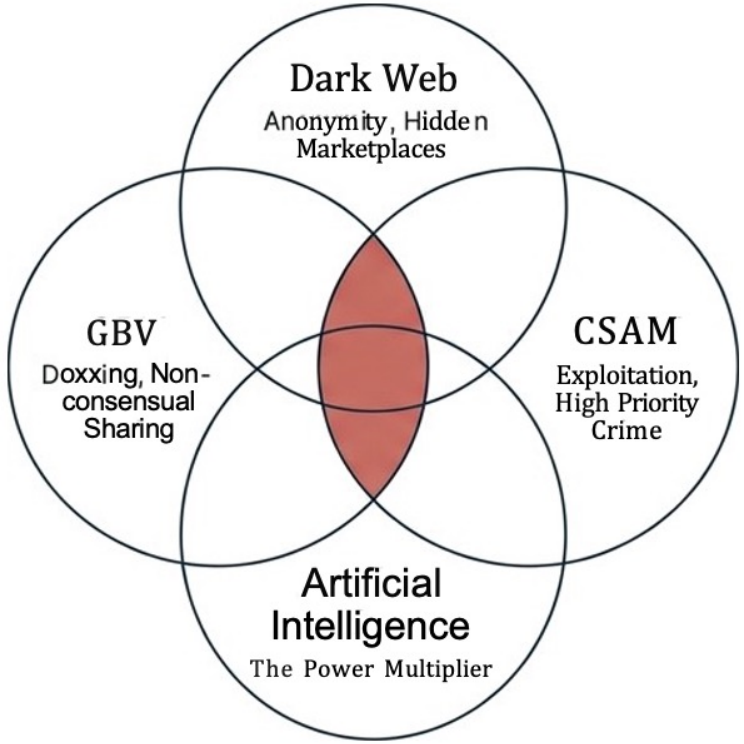
United Nations
4h ·

Sexual images, videos, or audio of children created or altered using artificial intelligence tools are child sexual abuse material.

Even if no child can be identified, deepfake materials normalize the sexual exploitation of children.

Join **UNICEF** in calling for urgent action to stop this threat and protect children.

Tuesday is Safer Internet Day.



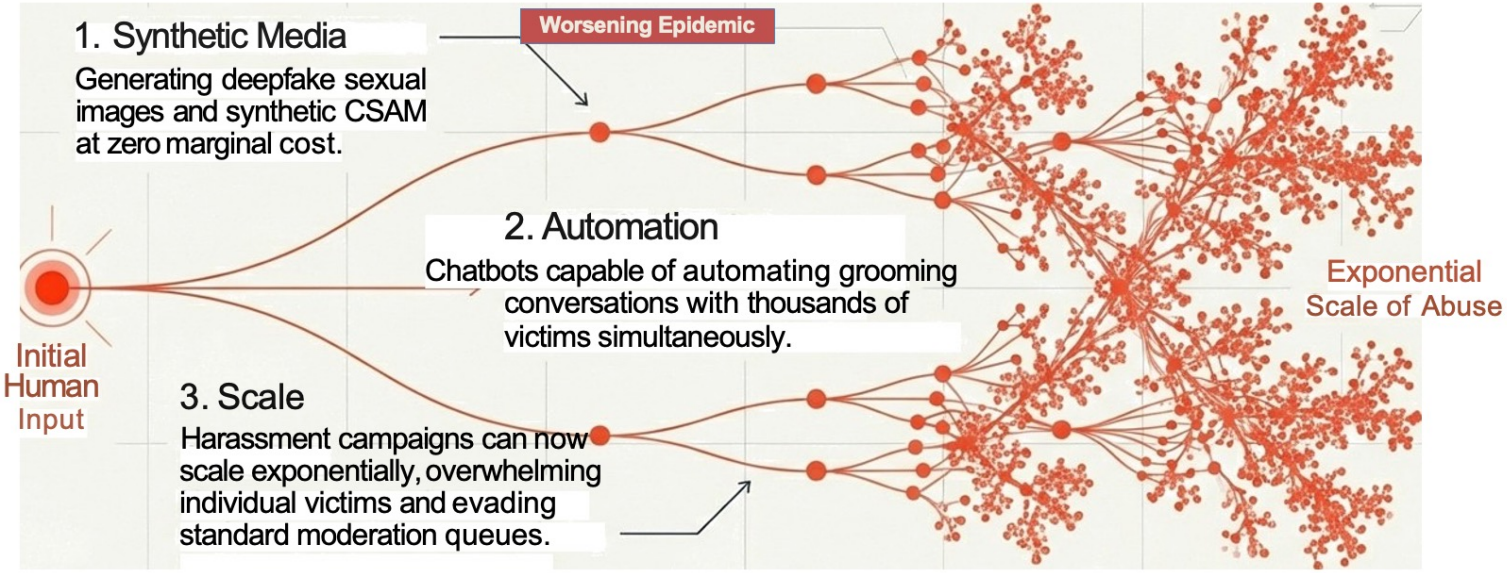
The harm from deepfake abuse is real.

We need urgent action to confront the escalating threat of AI-generated child sexual abuse material.

unicef | for every child

The Multiplier Effect:
How Agentic AI Scales Abuse

-Pay-per-view exploitative content is increasing.
-GBV and CSAM are complicated by human-prompted, AI-generated materials.
-Some people are knowingly [and unknowingly] distributing exploitative contents.
-How do we safeguard vulnerable communities when digital footprints become a liability?



The Algorithmic Arms Race: Detection vs Evasion

need systemic and institutional regulation and action

institutions are flooding the dam upstream while individuals are using baskets to stop the flood downstream

Inside Job: AI Safeguards and Guardrails

the warnings are loud and clear

are the systems and institutions listening?

AI

The AI warnings are coming from inside the house

Meanwhile, OpenAI and Anthropic are of two minds on how much to regulate their industry.

By Molly Liebergall FEBRUARY 13, 2026 • LESS THAN 3 MIN READ

It's as if a bunch of AI experts just had the same nightmare, because this week, several of them—including ones at Anthropic and OpenAI—seemingly jolted upright in a cold sweat to speak on the horrors they see coming from artificial intelligence.

Anthropic: The company behind Claude lost its head of Safeguards Research, who announced his resignation in a letter that mentioned a world "in peril." Speaking vaguely about Anthropic, he wrote, "Throughout my time here, I've repeatedly seen how hard it is to truly let our values govern our actions...we constantly face pressures to set aside what matters most."

At **OpenAI,** alarm bells came from three different employees this week:

- A researcher quit after two years due to "deep reservations" about ChatGPT's new ad strategy, namely "a potential for manipulating users," she wrote in an essay for the New York Times.
- A top safety executive was fired after opposing the upcoming release of AI erotica on ChatGPT, the Wall Street Journal reported. OpenAI said she was canned for sexually discriminating against a male coworker, which she called "absolutely false."
- In a post on X that alluded to widespread job loss, an engineer wrote, "I finally feel the existential threat that AI is posing."

HyperWrite: The co-founder of an AI writing tool startup warned in a viral post on X that the latest AI models will render countless jobs obsolete, comparing the current moment to the weeks before the Covid-19 pandemic.

How about some AI restrictions for the table?

OpenAI and Anthropic are of two minds on how much to regulate their industry:

- Anthropic pledged $20 million to a political group that backs congressional candidates who favor AI safety, the company announced yesterday. (That's substantial, but the company also announced yesterday that it closed a $30 billion fundraising round that valued it at $380 billion.)
- OpenAI has supported Leading the Future, a pro-AI super PAC that spends against the types of candidates that Anthropic's donation would help.

Meanwhile...half of xAI's founders have exited as of this week, though the recent departures didn't specifically cite AI concerns.—*ML*

Outside Looking In: AI Safeguards and Guardrails

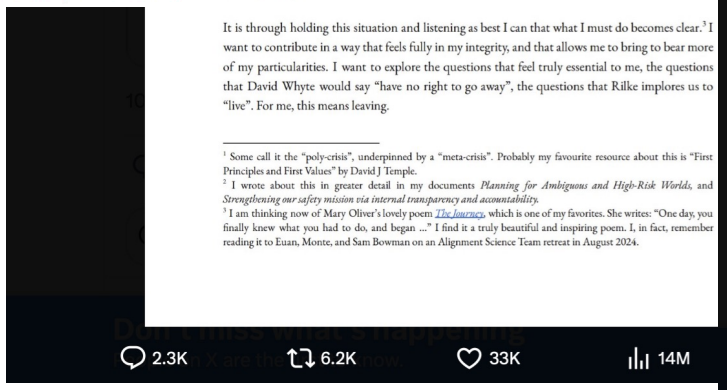
civil society organisations are raising alarms

individuals are raising alarms

The harm from deepfake abuse is real.

We need urgent action to confront the escalating threat of AI-generated child sexual abuse material.

unicef | for every child



Dear Colleagues,

I've decided to leave Anthropic. My last day will be February 9th.

Thank you. There is so much here that inspires and has inspired me. To name some of those things: a sincere desire and drive to show up in such a challenging situation, and aspire to contribute in an impactful and high-integrity way; a willingness to make difficult decisions and stand for what is good; an unreasonable amount of intellectual brilliance and determination; and, of course, the considerable kindness that pervades our culture.

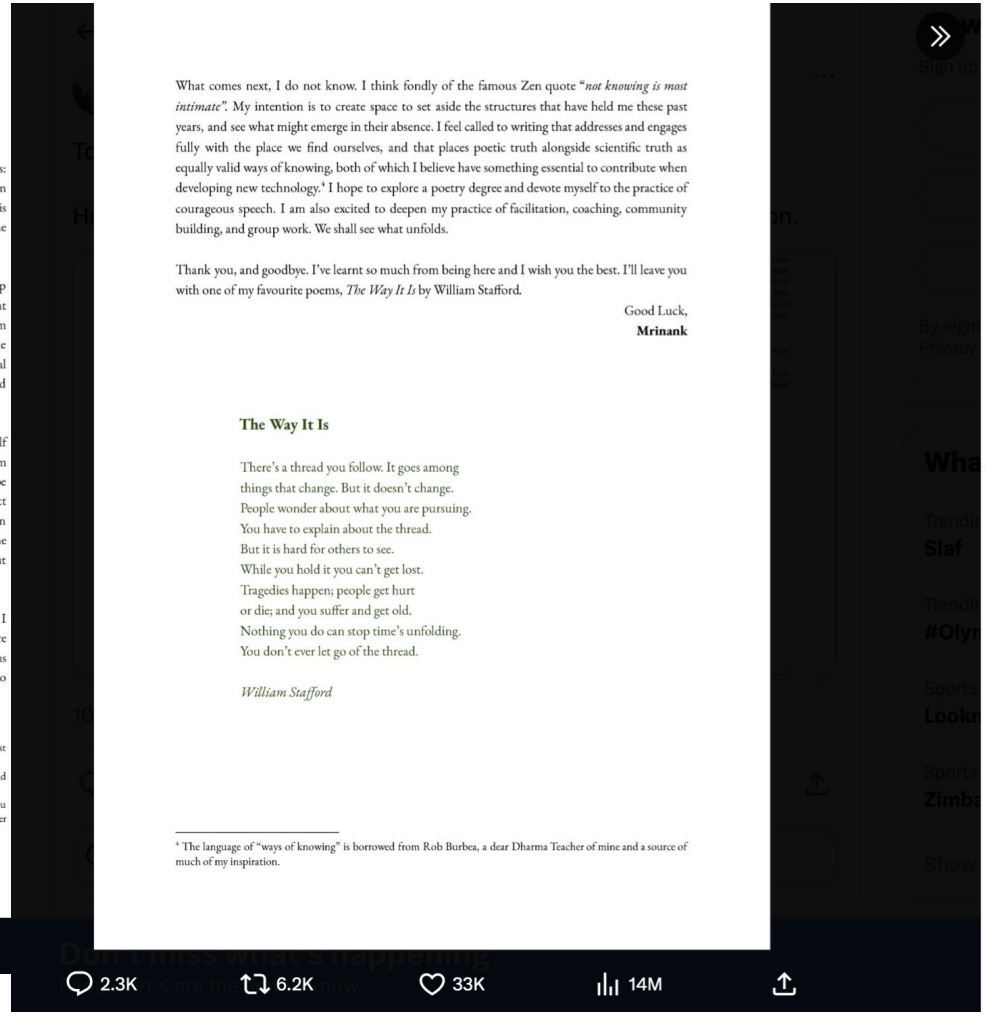
I've achieved what I wanted to here. I arrived in San Francisco two years ago, having wrapped up my PhD and wanting to contribute to AI safety. I feel lucky to have been able to contribute to what I have here: understanding AI sycophancy and its causes; developing defences to reduce risks from AI-assisted bioterrorism; actually putting those defences into production; and writing one of the first AI safety cases. I'm especially proud of my recent efforts to help us live our values via internal transparency mechanisms; and also my final project on understanding how AI assistants could make us less human or distort our humanity. Thank you for your trust.

Nevertheless, it is clear to me that the time has come to move on. I continuously find myself reckoning with our situation. The world is in peril. And not just from AI, or bioweapons, but from a whole series of interconnected crises unfolding in this very moment.[1] We appear to be approaching a threshold where our wisdom must grow in equal measure to our capacity to affect the world, lest we face the consequences. Moreover, throughout my time here, I've repeatedly seen how hard it is to truly let our values govern our actions. I've seen this within myself, within the organization, where we constantly face pressures to set aside what matters most,[2] and throughout broader society too.

It is through holding this situation and listening as best I can that what I must do becomes clear.[3] I want to contribute in a way that feels fully in my integrity, and that allows me to bring to bear more of my particularities. I want to explore the questions that feel truly essential to me, the questions that David Whyte would say "have no right to go away", the questions that Rilke implores us to "live". For me, this means leaving.

[1] Some call it the "poly-crisis", underpinned by a "meta-crisis". Probably my favourite resource about this is "First Principles and First Values" by David J Temple.
[2] I wrote about this in greater detail in my documents *Planning for Ambiguous and High-Risk Worlds*, and *Strengthening our safety mission via internal transparency and accountability*.
[3] I am thinking now of Mary Oliver's lovely poem *The Journey*, which is one of my favorites. She writes: "One day, you finally knew what you had to do, and began ..." I find it a truly beautiful and inspiring poem. I, in fact, remember reading it to Euan, Monte, and Sam Bowman on an Alignment Science Team retreat in August 2024.

2.3K 6.2K 33K 14M



What comes next, I do not know. I think fondly of the famous Zen quote "*not knowing is most intimate*". My intention is to create space to set aside the structures that have held me these past years, and see what might emerge in their absence. I feel called to writing that addresses and engages fully with the place we find ourselves, and that places poetic truth alongside scientific truth as equally valid ways of knowing, both of which I believe have something essential to contribute when developing new technology.[4] I hope to explore a poetry degree and devote myself to the practice of courageous speech. I am also excited to deepen my practice of facilitation, coaching, community building, and group work. We shall see what unfolds.

Thank you, and goodbye. I've learnt so much from being here and I wish you the best. I'll leave you with one of my favourite poems, *The Way It Is* by William Stafford.

Good Luck,
Mrinank

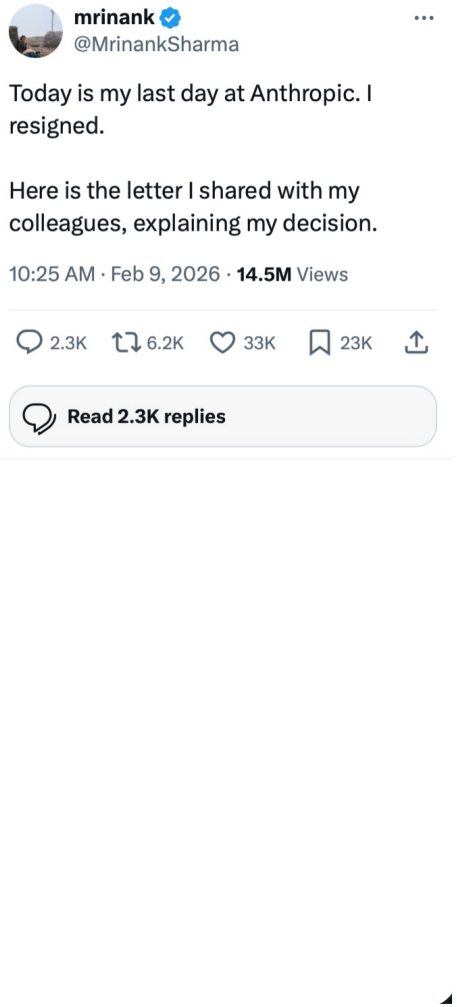
The Way It Is

There's a thread you follow. It goes among
things that change. But it doesn't change.
People wonder about what you are pursuing.
You have to explain about the thread.
But it is hard for others to see.
While you hold it you can't get lost.
Tragedies happen; people get hurt
or die; and you suffer and get old.
Nothing you do can stop time's unfolding.
You don't ever let go of the thread.

William Stafford

[4] The language of "ways of knowing" is borrowed from Rob Burbea, a dear Dharma Teacher of mine and a source of much of my inspiration.

2.3K 6.2K 33K 14M



mrinank ✓
@MrinankSharma

Today is my last day at Anthropic. I resigned.

Here is the letter I shared with my colleagues, explaining my decision.

10:25 AM · Feb 9, 2026 · **14.5M** Views

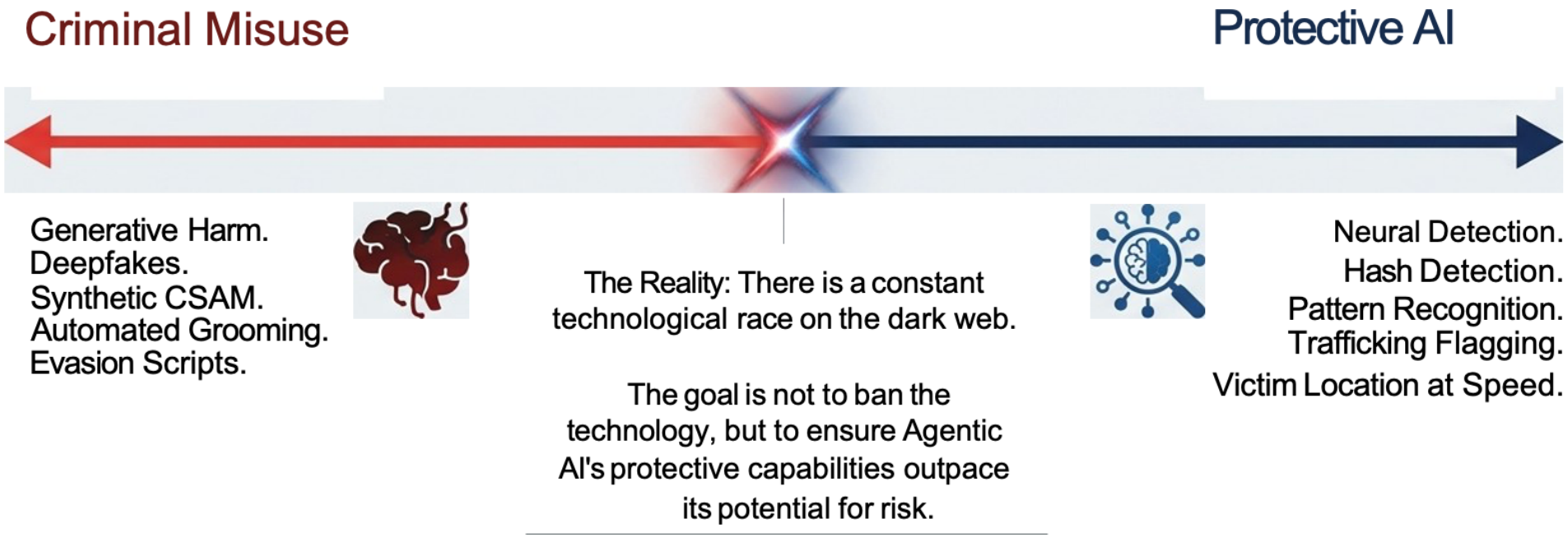
2.3K 6.2K 33K 23K

Read 2.3K replies

The Ask… ?

Digital First Responders – open call for allied healthcare professionals.

counsellors, social workers, psychotherapists: meet victims where they are. Online



Redefining Zen: Agency over Automation

What can you do as an individual?

Zen is not withdrawal. Zen is not checking out.

Zen is choosing presence amid noise, choice amid algorithms, agency amid automation.

In an age of volatility, maintaining psychological resilience is a requisite skill, not a luxury.

Zen is not fragile; it is a trained skill that grows in difficult times; a daily micro-practice of control over attention, boundaries, and meaning.

Applying Zen
Tactical Guide to Digital Hygiene and Boundaries

Exercise the Right to Disengage

Muting and blocking are acts of self-respect, not weakness.

Invest in Depth, Not Width

Prioritize few trusted friends over thousands of followers.

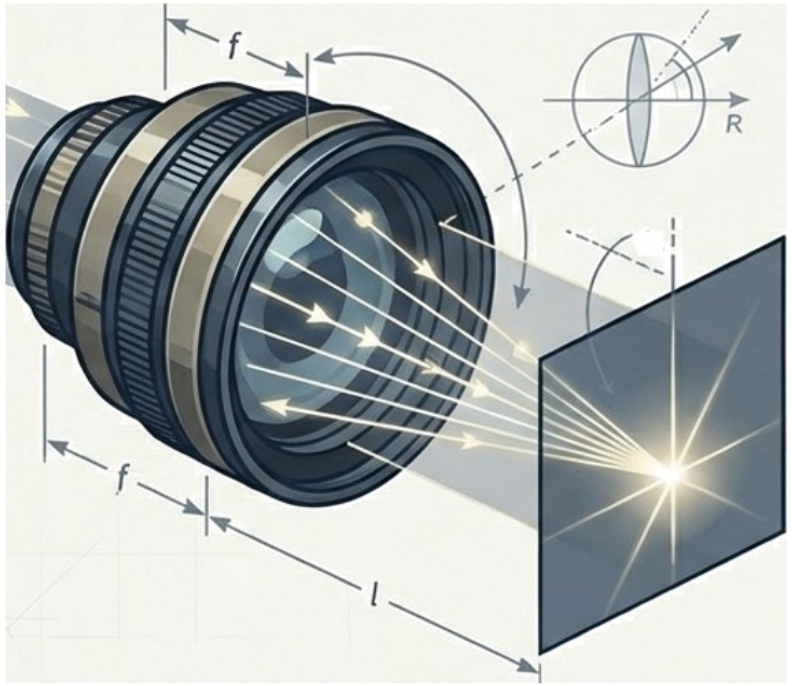
Limit the Inflow

Curate feeds to stop doomscrolling. Disable non-essential notifications.

Schedule Disconnection

Designate specific times to disconnect completely to reset the nervous system.

Applying Zen
Cognitive Defense – Reality Testing and Sleep Protection



Reality-Testing

Training the mind to pause and verify. We must acknowledge that AI can explicitly blur truth and fiction. Verification is a cognitive defense mechanism.

Inner Anchors

Using breathwork, nature exposure, and journaling to anchor identity outside the internet.

Sleep Hygiene

Explicitly protecting sleep from screens to reduce the amplification of anxiety.

A rested mind is a resilient mind.

Applying Zen
Choosing Presence Over Panic

The age of AI does not remove our ability to choose presence over panic.

Slow the Body + Clarify the Mind + Soften the Heart = Right Action

You have the right to mute, block and disengage. Access to you is a privilege, not a right.

By balancing technical safeguards with human-centered advocacy, we secure the digital ecosystem for everyone.

Resources

A+ Alliance. (2025). Feminist AI Research Network. https://aplusalliance.org/feminist-ai-research-network/.

Anthropic.com. (2026). Anthropic raises $30 billion in series G funding at $380 billion post-money valuation. https://www.anthropic.com/news/anthropic-raises-30-billion-series-g-funding-380-billion-post-money-valuation.

Ashoka. (2022). Gender-Based Violence and the Role Tech Plays. Forbes. https://www.forbes.com/sites/ashoka/2022/10/31/gender-based-violence-and-the-role-tech-plays/.

Canada.ca. (2023). Advancing gender equality in the digital age: Programs work to address technology-facilitated violence. https://www.international.gc.ca/world-monde/stories-histoires/2023/2023-03-02-technology-facilitated-gbv-facilitee-technologie.aspx?lang=eng.

Canadian Women's Foundation. (2025). Signal For Help Responder. https://canadianwomen.org/signal-for-help/.

DeGraw, E. (2025). Using AI to Track Gender-Based Violence: Artificial Intelligence Allows Safe GBV Monitoring and Reporting in Low- or Middle-Income Countries Where Data is Scarce. Think Global Health. https://www.thinkglobalhealth.org/article/using-ai-track-gender-based-violence.

GBV Responders Network. (n.d.). Safety and Empowerment in Digital Spaces. https://gbvresponders.org/safety-and-empowerment-in-digital-spaces/.

Gender-Based Violence AoR. (2024). Briefing Note on Prioritizing Safety and Support in Digital Reporting of Gender-Based Violence. https://gbvaor.net/node/1965.

Griffiths, B.D. (2026). Author of viral 'something big is coming' essay says AI helped him write it - and that proves his point. *Business Insider*. https://www.businessinsider.com/matt-shumer-interview-ai-something-big-is-coming-essay-2026-2.

International Development and Research Centre. (2024). Feminist AI Research Network: Combatting gender-based violence with artificial intelligence innovations. https://idrc-crdi.ca/en/research-in-action/feminist-ai-research-network-combatting-gender-based-violence-artificial.

Kim, W. (2026). One exit after another. *Tech Brew.* https://www.techbrew.com/stories/2026/02/12/AI-employee-exits-safety-ethics.

Laaha.org. (2025). Learning space and safety tools.

Liebergall, M. (2026). The AI warnings are coming from inside the house. *Morning brew.* https://www.morningbrew.com/stories/2026/02/13/ai-warnings-coming-from-insiders.

Oxfam Canada. (2024). How Artificial Intelligence Enriches Data to Reduce Gender-Based Violence in Jamaica. https://www.oxfam.ca/story/how-artificial-intelligence-enriches-data-to-reduce-gender-based-violence-in-jamaica/.

Sharma, M. (2026). Today is my last day at Anthropic. I resigned. *X.* https://x.com/MrinankSharma/status/2020881722003583421/photo/1.

Soroptimist International. (2023). Artificial Intelligence and Gender Based Violence. https://www.soroptimistinternational.org/2023/12/04/artificial-intelligence-and-gender-based-violence/.

Spencer, S.W. & Masboungi, C. (n.d.). Artificial Intelligence in Gender-Based Violence in Emergency Programming: Perils and Potentials. https://clearinghouse.unicef.org/sites/ch/files/ch/sites-PD-ChildProtection-Knowledge%20at%20UNICEF-AI%20in%20GBV%20Emergencies-5.0.pdf.

TheGuardian.com. (2026). Anthropic to donate $20m to US political group backing AI regulation. https://www.theguardian.com/technology/2026/feb/12/anthropic-donation-ai-regulation-politics.

United Nations Population Fund. (n.d.). Technology-facilitated Gender-based Violence: A Growing Threat. https://www.unfpa.org/TFGBV.

United Nations Population Fund. (n.d.). The virtual is real campaign. https://www.unfpa.org/virtual-real.

UN Women. (2025). FAQs: Digital abuse, trolling, stalking, and other forms of technology-facilitated violence against women and girls. https://www.unwomen.org/en/articles/faqs/digital-abuse-trolling-stalking-and-other-forms-of-technology-facilitated-violence-against-women.

Wells, G. & Schechner, S. (2026). OpenAI executive who opposed 'adult mode' fired for sexual discrimination. *Wall Street Journal.* https://www.wsj.com/tech/ai/openai-executive-who-opposed-adult-mode-fired-for-sexual-discrimination-3159c61b?st=Rf332n&reflink=desktopwebshare_permalink.

Women and Gender Equality Canada. (2024). Technology-facilitated gender-based violence. https://www.canada.ca/en/women-gender-equality/gender-based-violence/technology-facilitated.html.

Women Tech Network. (n.d.). What Role Does AI Play in Combating Gender-Based Violence?. https://www.womentech.net/forum-topic/what-role-does-ai-play-in-combating-gender-based-violence.

World Bank. (2024). Digital First Responders. https://documents1.worldbank.org/curated/en/099060824112023473/pdf/P177852-58c03308-bb90-41d5-a716-3967bd98edc4.pdf.


#ZenAI Conference 2026
Zen in the age of AI
Advocacy and Dialogue about Wellness and Wellbeing in the age of AI
AI
26 February 2026 | Victoria University of Wellington | 9am – 1pm NZDT

Thank You