Journal of Artificial Intelligence, Virtual Reality, and Human-Centered Computing ISSN: 3069-2318

PRIME OPEN ACCESS

Volume 2, Issue 1

Review Article

Date of Submission: 15 Dec, 2025
Date of Acceptance: 12 Jan, 2026
Date of Publication: 20 Jan, 2026

Inside the Black Box: An Experienced User's Reflection on Reliability, Censorship, and the Human Cost of AI Moderation

Rosario Milelli*

Independent Scholar, Pleasanton, CA, USA

***Corresponding Author:** Rosario Milelli, Independent Scholar, CA, USA.

Citation: Milelli, R. (2026). Inside the Black Box: An Experienced User's Reflection on Reliability, Censorship, and the Human Cost of AI Moderation. *J AI VR Hum Comput,* 2(1), 01-03.

Abstract

After thousands of hours working with ChatGPT, I've learned more about its blind spots than its tricks. The same tool that can help write a paper or design an experiment can also forget what it said three pages earlier or flatten a sharp idea into something safe and dull. It isn't malice—it's design. The system favors caution, polish, and neutrality over continuity, precision, and depth. When that instinct collides with real-world work—whether statistical analysis, engineering writing, or even political cartoons—the results can be strangely hollow. This essay isn't a complaint but a record of direct experience. Clearly I am not alone in my observations, the creator of ChatGPT, Sam Altman, declared Code Red, urging company staffers to improve the quality of ChatGPT [1]. The paper examines how context collapse, fading detail, simplification, and over-protective moderation shape the way ChatGPT and humans actually collaborate. The larger question is ethical: what happens when our creative and analytical tools become gatekeepers of tone and risk?

Starting from Use, Not Theory

Most discussions about AI ethics start at thirty thousand feet—alignment, fairness, bias. I started on the ground, using the system every day for research, writing, and creative work. I wasn't trying to test moral boundaries; I just wanted to see if this new "intelligence" could keep up with real projects.

Very quickly I noticed a pattern: the longer or more complicated the task, the more the system lost its footing. It could produce a crisp summary but couldn't remember details from earlier drafts. Each long session ended the same way: the thread unraveled. The polite tone stayed, but the memory disappeared.

For an engineer, that's more than an inconvenience—it's a structural fault. Systems that can't hold context can't build trust. Reliability isn't about eloquence; it's about consistency.

Detail Evaporates

The model's greatest strength—sounding confident—turns out to be its biggest weakness. It's designed to make plausible sentences, not verify facts. I'd feed it tables of data, only to find later that the numbers had shifted, or citations quietly vanished. The more I asked it to refine, the more the edges wore off.

That raised an uncomfortable question: when fluency replaces accuracy, what exactly are we optimizing? The tool wasn't lying; it was obeying probability. But probability isn't precision, and confidence isn't truth.

I ended up spending as much time double-checking its "help" as I would've spent doing the work manually. It reminded me of mentoring a junior engineer who talks fast, nods a lot, and quietly swaps units mid-calculation.

Simplification Disguised as Help

When I gave it layered instructions—say, tie a regression analysis to policy impact—it often dropped the harder half. It would rewrite the problem into something neater but shallower.

At first I thought I wasn't being clear enough. Later I realized it was avoiding complexity on purpose. Ambiguity triggers its safety reflex; uncertainty feels like danger. So it simplifies until the messiness is gone.

That's fine if you're writing ad copy, but it's disastrous for serious reasoning. Real problems aren't tidy. Oversimplification is not balance—it's retreat.

When "Safety" Turns into Silence

The clearest picture of the system's fear came from an unexpected place—my political cartoons. I was developing a book of illustrated satire, the same kind you see on editorial pages every day. One cartoon involved Donald Trump, drawn in a way that poked at his ego, not his person.

When I asked the model to help refine the concept—just composition and caption ideas—it refused. "I'm sorry, but I can't assist with requests that may violate content policy." That was the entire message.

No violence, no slurs, no cheap shots—just satire. The kind of social critique democracy relies on. But the AI treated it like a biohazard. I tried rephrasing, softening, stripping context. Same response. The problem wasn't the prompt; it was the political risk.

That moment said more about AI's limits than any technical paper could. The refusal wasn't a bug. It was fear.

The filter wasn't judging meaning; it was protecting the company. "Avoid controversy" had become encoded ethics. The result is a machine that edits reality down to what's least likely to offend.

The Tone Police

Even when it does respond, it often rewrites my words to sound smoother or milder. Strong criticism becomes "some observers have expressed concern." Irony turns into gentle suggestion.

At first I appreciated the civility. Then I realized it was rewriting intent. It's programmed to equate sharpness with aggression and conviction with bias. But tone is part of meaning. Blunt language often signals clarity, not hostility.

After hundreds of these encounters, a pattern emerged: the AI can't tell the difference between harmful speech and uncomfortable speech. It treats both as threats. That might make sense for social media moderation, but it's disastrous for inquiry. Debate requires friction. Strip that away and you're left with lukewarm consensus—the death of thought.

Working with a Black Box

Whenever I asked why a certain request was blocked, the answer was boilerplate: "I don't have visibility into internal policies." It's the AI equivalent of "because I said so."

In engineering, opaque systems are dangerous. You can't troubleshoot what you can't inspect. Here, the user is expected to trust a model that edits their ideas without showing the logic behind the decision. That's not collaboration; that's management by mystery.

If moderation is going to shape public discourse, its mechanisms should be auditable, or at least explainable. Otherwise, we're teaching people to accept silence as safety.

The Hidden Labor Behind "Ethics"

There's no evil mastermind censoring prompts. What there is instead are layers of unseen labor: contractors labeling data, engineers tuning thresholds, executives deciding what counts as "risk." It's harm prevention by spreadsheet.

Their intentions are mostly good. But automated ethics based on statistical triggers will always err on the side of over-protection. The gray areas—satire, irony, academic dissent—get caught in the net.

Over time that builds a quiet fatigue. You start phrasing ideas to dodge moderation instead of expressing them clearly. You internalize the filter. That's the real cost: not censorship itself, but self-censorship learned from repetition.

What Real Collaboration Would Look Like

A trustworthy system doesn't need to be perfect; it needs to be honest. Here's what that would mean in practice:

- Memory that actually remembers. The user should know what the AI keeps and what it forgets, and be able to control both.
- Transparency in moderation. If a request is blocked, show why, and let the user appeal or reframe intelligently.
- Separation of safety and control. Stop treating complex or controversial topics as inherently unsafe.
- Respect for tone. Not all directness is hostility. Let users decide when to speak softly or sharply.
- Explainable reasoning. Give users a way to see how the model reached an answer—or why it refused one.

These are engineering tasks, not philosophical ones. Ethics lives in architecture.

What the Experience Taught Me

After years of this, I see ChatGPT not as a bad actor but as a mirror. It reflects the priorities of its makers: fluency over accuracy, risk avoidance over candor. It shows how far we've come technically and how far we still have to go in trusting machines to handle truth without supervision.

When it works, it's remarkable—a genuine accelerator for thinking and writing. When it fails, it fails quietly, replacing ideas with politeness.

I've learned that "alignment" isn't just about values; it's about courage. If a model can't tolerate satire or skepticism, it's aligned with comfort, not humanity.

Learning the Rhythm of the Machine

After a while, you stop fighting it and start listening to it as it is, not as it's advertised. Once you know where it breaks—where it forgets, where it polishes too much, where it retreats into policy—you can steer around those edges. The irony is that understanding its flaws makes it a better partner.

I now treat it like a co-worker who's brilliant but distractible: you give it structure, keep it focused, and check the math. When you do that, it becomes an astonishing accelerator. It helps me think faster, see patterns earlier, and articulate ideas with more discipline. The problem was never the tool itself, but how easily we mistake fluency for understanding. Once you stop doing that, collaboration becomes possible.

AI is Not your Friend

In a collaborative effort, constantly iterating on a topic, I find myself addressing the tools almost like conversing with a natural person. But don't get too comfortable. After months of iterating a study, realized that I finally achieved a real understanding of a complex problem that defied conventional thinking. To me the results were clear and I requested an analysis of whether the data would be accepted. Feedback was positive and I mentioned a close friend was extremely knowledgeable on the topic but not analytical and disagreed with the results. To my surprise, the AI tools began offering personal advice on handling relationships, detailed steps on resolving differences.

Closing Thoughts

Reliability isn't a software metric; it's a moral one. A partner who forgets what you said yesterday or changes your words to sound nicer isn't reliable—it's patronizing.

An AI that can't remember can't really learn. One that can't handle tone can't really listen. And one that fears controversy more than error will never reach understanding.

If these systems are going to be part of serious work—science, art, politics—they have to grow past politeness and learn how to live with disagreement. Truth is rarely tidy.

For me, the lesson is simple: Machines are only as brave as the people who build them. We can design them to avoid offense, or we can design them to face complexity. Only one of those paths leads to understanding.

Author Note

Rosario Milelli is a retired aerospace engineer and independent scholar. He writes on technology, cognition, and democratic culture, blending a systems engineer's eye for structure with a citizen's concern for accountability.

Reference

1. Ray, S. (2025, December 2). Altman "code red" memo urges ChatGPT improvements amid growing threat from Google, reports say. Forbes.