AI & SOCIETY
https://doi.org/10.1007/s00146-025-02746-4

CURMUDGEON CORNER

The algorithm that hates for our own good

Dana Mahr[1]

Received: 27 October 2025 / Accepted: 11 November 2025

Artificial intelligence has become the world's most zealous hall monitor. Always watching, seldom listening. It deletes before it thinks, applauds itself for efficiency, and claims neutrality while performing the oldest social trick in the book: policing the margins. Today's "responsible AI" systems, armed with billions of dollars in ethics budgets and fairness dashboards, have turned online moderation into a mechanized ritual of moral hygiene. But behind this civility theater lies a disquieting truth: the machines designed to fight hate increasingly learn to hate on command.

Every few seconds an algorithm somewhere deletes a post, bans a user, or buries a conversation. The stated goal is noble—protecting people from harm. The actual effect is often perverse. A Black activist quoting a slur in protest finds her words flagged as abusive, while an image model quietly manufactures new stereotypes in meme form. A queer forum disappears under "safety filters," but misogynistic vitriol slides through because it is phrased with enough irony. The same architectures that remove hate speech now generate it with astonishing fluency. In the age of large language models, our censors and our provocateurs share the same silicon lineage.

This is not an accident. It is governance by feedback loop. AI learns from the world as it is, and the world rewards outrage. The result is a self-reinforcing moral economy of visibility: rage sells, moderation silences, and both run on the same infrastructure. What we call "harm reduction" too often looks like the industrial outsourcing of judgment to code. And when the code errs, it errs in ways that echo centuries of hierarchy. The activist is punished; the troll goes viral. The voices that most need protection are the ones most likely to be mistaken for threats.

We are told that the solution lies in better data, more diverse annotators, and "fairness-by-design." But these slogans confuse repair with reform. A fairness pipeline is still a pipeline. The deeper problem is that commercial moderation operates under corporate secrecy. The public is asked to trust systems it cannot inspect, enforcing rules it cannot contest, in digital spaces that now double as our public sphere. "Community standards" sound benign until one realizes that the community in question is a trillion-dollar platform whose shareholders define civility.

The paradox is moral as much as technical. Each new generation of moderation AI arrives wrapped in benevolence: protecting users, preventing hate, preserving democracy. Yet each iteration expands the scope of automated control. The machine that once flagged slurs now evaluates tone, emotion, even intent. "Contextual AI" is sold as sensitivity; in practice, it is surveillance with a velvet touch. We are approaching a world where dissent can be categorized as toxicity, and where the "trust and safety" industry determines not only what can be said, but what can be *meant.*

To be clear: the problem is not that machines moderate poorly, but that we have mistaken moderation for governance. Real governance requires accountability. Someone who answers when things go wrong. In the algorithmic regime, responsibility dissolves in layers of technical abstraction. A user appeals a decision only to receive a friendly automated apology: "Our systems sometimes make mistakes." Who, exactly, is "our"? A contractor in Manila? A developer in California? A training dataset scraped from Reddit? Accountability vanishes into distributed complexity, a perfect bureaucratic shield disguised as innovation.

There is an old joke in cybernetics that every system eventually becomes a mirror. AI moderation has reached that stage: it reflects our collective discomfort with conflict, our managerial desire for frictionless order. The tragedy is that the very messiness being sanitized is what democracy requires. Public discourse cannot be both safe and free; it must remain dangerous enough to disturb. In cleaning the surface of the internet, we risk sterilizing the soil of dissent.

Perhaps this is why the rhetoric of "responsible AI" rings hollow. Responsibility has been redefined as compliance. A model is considered "ethical" once it passes a checklist.

✉ Dana Mahr
dana.mahr@kit.edu

[1] Institute for Technology Assessment and Systems Analysis, Karlsruhe Institute of Technology, Karlsruhe, Germany

Published online: 28 November 2025

Transparency means a glossy blog post; oversight means a self-published audit. Meanwhile, the underlying political economy (one that trades human expression for ad revenue) remains untouched. We celebrate AI ethics as though it were a moral achievement, when it is merely risk management by other means.

The truly uncomfortable question is not how to make moderation fairer, but whether certain forms of automated moderation should exist at all. Do we really want an opaque machine adjudicating the boundaries of hate, irony, or political anger? The defenders will reply that scale demands it: billions of posts, millions of users, too much for human review. But scale is not a law of nature; it is a business choice. We built platforms too large for human governance, then declared the machines indispensable because the humans no longer fit. This is less technological necessity than institutional abdication.

A more honest vision of responsibility would begin with humility. Admit that no model can define hate outside of history. Accept that some errors are moral, not statistical. Reintroduce humans, not as janitors cleaning up after AI, but as decision-makers whose judgment counts. Create public, inspectable mechanisms for appeal and redress. And most importantly, separate the policing of speech from the profit motives of those who benefit from engagement. The internet does not need smarter filters; it needs democratic ones.

The irony, of course, is that AI could help us build those very infrastructures—if we stopped treating it as a substitute for politics. Imagine using machine learning to map moderation disparities across demographics, or to simulate the social consequences of different policy choices before implementation. That would be genuine "responsible AI": not the automation of morality, but the augmentation of collective self-governance. Instead, we have settled for predictive censorship wrapped in ethical branding.

What is at stake here is not just online civility, but the shape of public reason itself. Every generation invents its own form of silence. The Victorians had decorum; the Cold War had propaganda; we have content moderation. Our age's genius is to automate the silence so thoroughly that no one notices. Posts vanish quietly. Accounts shadow-ban themselves into oblivion. And the platforms proclaim success: "The system is learning." Indeed it is. But from whom, and to what end?

The curmudgeon in me suspects we have mislearned the lesson of AI moderation. The goal was never to teach machines to distinguish hate from critique; it was to teach humans to stop asking uncomfortable questions about power. The algorithm is merely the polite face of governance without governors: a mask for political choices recast as engineering challenges.

So here is a modest proposal: let us retire the phrase "responsible AI moderation." Call it what it is (automated governance) and treat it accordingly. Subject it to the same public scrutiny we demand of courts, legislatures, and media institutions. If AI now defines the boundaries of acceptable speech, then it should also be answerable to those it governs. Until then, the promise of "safe" online spaces will remain what it has always been: a euphemism for someone else deciding what the rest of us are allowed to say.

Yours sincerely,

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Author contribution D.M. conceived, researched, and wrote the manuscript in its entirety. D.M. is solely responsible for the content and its final version.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.