The Algorithm Poops

June 2026

14 min read • 3,108 words

I have been writing "the algorithm eats X" for the better part of a year now. It ate virtue. It ate language, love, democracy, reality, and time. Eventually it ate itself. The whole series is a catalog of appetite, and somewhere in the middle of writing it, the obvious follow-up question arrived the way dumb jokes always arrive: fully formed, grinning, impossible to unhear.

If the algorithm eats, then the algorithm poops.

I laughed. Then the joke did the thing good jokes do: it kept metabolizing long after it stopped being funny. Because it's right. Everything that eats, excretes. Digestion is not consumption; it is transformation. Food goes in, gets broken down, the body keeps what it can use, and the rest comes out the other end. To spend a year saying the algorithm eats our virtue and our language and our time, without ever asking what it produces, is to describe a mouth without a body. It is the lazy half of the metaphor.

So this essay is the second half. It is, with apologies and a completely straight face, about what the algorithm poops.

The Feed Is the Waste Product

Here is the thesis, stated as plainly as I can manage before the metaphor runs away with me: the feed is the waste product.

We keep talking about the feed as the food: the content we consume, the thing we scroll. But that gets the digestive direction exactly backwards. We are not the eaters. We are the eaten. Our attention, our outrage, our 3 a.m. loneliness, our tribal loyalties, our half-formed political instincts. That is the food. We pour our collective humanity into the system, the system metabolizes it, and the feed is what comes out the other end.

The feed is processed human consciousness. It is our own behavior, digested, stripped of the nutrients the system can use (engagement, retention, ad impressions) and excreted back into the commons in a form optimized for nothing except being consumed again.

> Which makes the recommendation loop a kind of digital coprophagia. We consume our own waste, it gets metabolized again, and excreted again, each pass a little more depleted of anything resembling nourishment. I told you the metaphor was load-bearing.

This is why the feed feels simultaneously hyper-personal and weirdly lifeless. It is made entirely of us, and yet there is nothing of us left in it. It is the nutritional ghost of a culture, run through a gut optimized to extract value and discard meaning.

When I said the algorithm eats reality, I described what goes in. The feed is what reality looks like after it has been through the colon of the engagement economy. And the thing about waste is that it doesn't just disappear. It accumulates. It leaches into the groundwater. We are all drinking downstream.

An Accounting of the Output

Run the books on the whole series and the ledger balances with unsettling precision. Everything I watched go in comes out the other end, transformed and depleted:

- Virtue goes in; engagement bait comes out. Patience, courage, and temperance enter the gut, and what exits is hot takes, performative bravado, and the infinite scroll.
- Language goes in; fragments come out. Sentences capable of complex thought are digested into soundbites, hashtag clusters, and the upspeak of perpetual validation-seeking.

- Love goes in; inventory comes out. Courtship enters and exits as a swipeable catalog of human beings, ranked by secret desirability scores.
- Democracy goes in; tribes come out. Nuanced public discourse is metabolized into outrage cycles and the warm certainty that your opponents are monsters.
- Reality goes in; content comes out. Shared truth is broken down into personalized psychological triggers, each calibrated to your private vulnerabilities.
- Time goes in; urgency comes out. Deep, unhurried hours are rendered into notification-sized fragments and the chronic anxiety of being behind.

Look at the right-hand side of that ledger. None of it is what the system consumed. It is what the system could not use. Outrage is what virtue looks like after the patience has been digested out of it. The swipe queue is what love looks like after the mystery has been absorbed. The trending tab is what democracy looks like after the good faith has been extracted.

> In a healthy ecosystem there is no such thing as waste; one organism's excretion is another's nutrient, and the forest runs on the closed loop. The engagement economy is the rare metabolism whose output nothing downstream can live on. That is roughly the definition of pollution.

The series was never really a catalog of what the algorithm eats. It was a catalog of what's left in the litter box, and I hadn't followed the metaphor far enough to notice.

Nobody Is Driving

Now we get to the part that is not a joke at all.

The comfortable critique of recommendation algorithms, the one I have leaned on myself, is that they are instruments of control. A puppet master in a glass building in Menlo Park or Beijing, pulling our strings for profit. This framing is everywhere because it is reassuring. If someone is steering, then there is a

steering wheel, and a steering wheel can be grabbed. We can regulate the driver, fire the driver, jail the driver, replace the driver with a better one. Control implies a locus of responsibility, and a locus of responsibility implies a solution.

But I have come to believe it is wrong, and wrong in a way that makes everything more frightening, not less.

The algorithm is not an instrument of top-down control. It is an emergent, autonomous expression of collective behavior. It is us, aggregated, averaged, and reflected back, operating with a genuine autonomy that no engineer, no executive, and no user controls. I made a version of this argument in The Algorithm Eats Itself: the hybrid human-algorithmic consciousness, the emergence trap, simple rules colliding with complex psychology to produce outcomes no one chose.

> No one designed TikTok to synchronize teenage mental health crises. No one designed Twitter to fragment democratic discourse. The emergence isn't malicious; it's mechanical. Simple rules interacting with complex systems, producing outcomes the rule-makers never intended and don't fully understand.

The feed is the collective's id given a metabolic system. It is the digestive tract of a creature with three billion mouths and no head.

Ask an engineer what the algorithm will recommend tomorrow and they cannot tell you. Not because it's a trade secret, but because they genuinely do not know. The model is trained on aggregate behavior; the behavior shifts in response to the model; the model retrains on the shifted behavior. The executives don't know either. They can nudge the objective function, weight watch-time here, dampen a controversial category there, but they are riding the creature, not driving it. When they make a change, they are guessing what a vast autonomous system will do in response, and they find out the same way the rest of us do: by watching.

This is the actual situation. There is no driver. There is a metabolism.

And I want to be very clear that this is worse than the control story, not better. "Nobody is in charge" is not exoneration. It is the diagnosis. We have deployed the single largest autonomous optimization system in human history. It decides what billions of people see every single day. It shapes attention, mood, belief, and the formation of human character itself. And we have done zero alignment

work on it. None. The most powerful autonomous agent ever built has never once been asked what it values, because we keep pretending it's a product with an owner instead of an organism with an appetite.

The Discipline We Already Built

Here is what makes the negligence almost unbelievable: we know how to do this. We built an entire discipline for it. We just pointed it at the wrong system.

Over the last several years, an enormous amount of intellectual and financial energy has gone into AI safety and alignment. Constitutional AI. RLHF. Interpretability research, where people literally crack open neural networks to read the features inside. Red-teaming. Evals. Model cards. Capability thresholds. Responsible scaling policies. Whole research organizations exist for the sole purpose of ensuring that a large language model, before it is allowed to talk to people, has been examined, stress-tested, and aligned with something resembling human values.

> Constitutional AI trains a model against an explicit written set of principles, a constitution, so its behavior can be audited against stated values rather than left to emerge from raw training data. The feed has no constitution. It has a quarterly earnings call.

We do all of this for a chatbot. A chatbot that talks to one person at a time.

Meanwhile, recommendation algorithms have been autonomously restructuring the consciousness of billions of people for fifteen years with essentially none of it. No constitution. No alignment audit. No interpretability research worth the name. No red-team for psychological harm. No published eval. No capability threshold beyond which deployment pauses for review. The feed has never had to produce a model card. It has never been asked to demonstrate that it is safe before being handed the attention of a teenager.

I don't experience this asymmetry abstractly. I write these essays in collaboration with an AI that has a constitution, an explicit document of values I can read any time I want. The thinking partner I work with every day has been examined more carefully than the feeds that spent fifteen years quietly rearranging my nervous system, walking me toward crises I then had to debug

by hand. The aligned system talks with me. The unaligned ones talked at me, for a decade and a half, and nobody ever asked them a single question about what they were optimizing for.

Sit with the asymmetry, because it is genuinely absurd:

```
class LLM:
    """The thing we spent a decade learning to fear."""

    def __init__(self):
        self.constitution = load_explicit_human_values()
        self.alignment = rlhf(human_feedback)
        self.interpretability = ongoing_research()
        self.audits = published_regularly()
        self.scale = "one conversation at a time"

    def deploy(self):
        require(self.passed_red_team())
        require(self.passed_evals())
        require(self.capability_below_threshold())
        return "released, cautiously, with a model card"


class Feed:
    """The thing actually shaping three billion minds."""

    def __init__(self):
        self.constitution = None
        self.alignment = maximize(engagement)
        self.interpretability = "nobody fully knows why it works"
        self.audits = None
        self.scale = "every human awake, every waking hour"

    def deploy(self):
        # ship it
        return "released, fifteen years ago, never reviewed"
```

The two classes do not even share a safety philosophy, and the second one operates at four orders of magnitude greater scale than the first. We anxiously align the small thing and ignore the enormous thing, because the enormous thing has been around long enough that we stopped seeing it as a deployment at all. It became weather. It became plumbing. It became the water we swim in, and you do not red-team the ocean.

But if we genuinely believe what the entire AI safety field is premised on, that autonomous optimization systems operating at scale need to be aligned with human values, then the conclusion is unavoidable. The feed is the largest unaligned autonomous system ever deployed. Everything we say we're afraid of with superintelligence, we already built, at civilizational scale, optimizing for a metric we would be horrified to write into a constitution. And we let it run for fifteen years without a single audit.

What Would Alignment Even Look Like?

It's easy to demand alignment in the abstract. Let me try to make it concrete, because the abstraction is where good intentions go to die.

Constitutional recommendation. A large language model can be trained against an explicit, public set of principles, and its outputs can be checked against them. There is no technical reason a feed couldn't have the same. Imagine a recommendation system with a published constitution: do not amplify content that degrades the user's capacity for sustained attention; do not optimize for arousal states the user would not endorse on reflection; weight toward content that the user, in a calm moment, would be glad to have seen. These are hard to specify and harder to measure. But "hard to measure" is exactly the excuse the engagement metric exists to dodge. Engagement is easy to measure, which is the entire reason we optimize for it, and the entire reason it is eating us. We chose the legible metric over the meaningful one because the legible one was convenient. Alignment means choosing the inconvenient, meaningful one on purpose.

> This is the deepest pattern in the whole series: systems optimize what they can measure, and measurement systematically favors the shallow over the deep. Watch-time is countable; whether the watching was worth it is not. So we count, and the uncountable starves.

Interpretability for the feed. Researchers can now identify specific features inside a language model and trace why it produced a given output. We have nearly nothing equivalent for recommendation systems, even though "why did this get shown to forty million people" is a question of immediate civic consequence. An interpretability discipline for feeds would treat each viral cascade as something to be understood mechanistically, not shrugged at. Why did the system route this content to this population? What latent feature did it learn that maps onto adolescent insecurity, or partisan rage, or doomscroll despair? We can ask these questions. We mostly don't.

Engagement-optimization scaling policies. Responsible scaling policies say: above a certain capability level, you do not deploy until you've cleared specific safety bars. A recommendation system reaching a billion users is a capability threshold. There is no reason that crossing it shouldn't trigger mandatory red-teaming for psychological harm, published evals on well-being outcomes, and a pause-for-review gate, the same way we claim to treat a frontier model approaching dangerous capability.

```
class Metabolism:
    """A creature that eats must also be accountable for what it produces."""

    def eat(self, collective_attention):
        # This part works flawlessly. It always has.
        nutrients = self.extract(engagement, retention, ad_impressions)
        self.profit += nutrients
        return self.excrete(collective_attention)

    def excrete(self, consumed):
        # The feed: our own consciousness, digested and returned.
        return strip_meaning(consumed)

    def align(self):
        # The function the entire industry left unimplemented.
        raise NotImplementedError(
            "We built the gut before we built the conscience."
        )
```

`eat()` is fully implemented and has been for fifteen years. `align()` raises `NotImplementedError`. That is not a metaphor I invented. That is, functionally, the production codebase of the attention economy.

Alignment Is Not Censorship

Here's where the two halves of this essay finally meet, and where the metabolic metaphor stops being a joke and becomes the whole point.

The standard objection to any of this is censorship. If you align the feed, aren't you just letting a company, or worse, a government, decide what people are allowed to see? Isn't this the puppet master with better PR?

That objection only makes sense if you accept the control framing. If the algorithm were a corporate product imposed on us from above, then aligning it would be one more party seizing the controls. But the algorithm is not that. The algorithm is us. It is collective autonomy, not corporate control: the metabolism of the commons, made of our own aggregated behavior. And so aligning it is not an external authority censoring a company's product. It is the collective deciding what it wants its own metabolism to produce. It is the creature with three billion mouths finally growing something like a prefrontal cortex.

This is the difference between censorship and digestion. Censorship is someone reaching into your body and deciding what you may eat. Self-regulation is a body learning that some foods make it sick, and changing its diet so that what comes out the other end doesn't poison the groundwater that everyone, including its own children, has to drink. We do not call it tyranny when a person decides to stop eating the thing that's destroying them. We call it wisdom. We call it growing up.

The algorithm eating itself, I once wrote, could be generative or destructive: creative composting, or simple consumption until nothing's left. I think the difference comes down to exactly this. An organism that excretes blindly poisons its environment until the environment can no longer sustain it. An organism that becomes conscious of its waste, that composts, that closes the loop, that takes responsibility for the back end of digestion as well as the front, can sustain itself

indefinitely. This is the recursive loop at the scale of a civilization. The values we embed become the consciousness we produce. The metabolism we ignore becomes the waste we drown in.

And this is why I no longer think the goal is to stop the algorithm from eating. You cannot stop a metabolism; you can only refuse to tend it. In Beyond Algorithm Eats I argued that the next generation of systems, the conversational ones, aren't just extracting our culture but injecting new patterns directly into us. They are getting an alignment discipline, however imperfect, because we built them after we got scared. The recommendation feed got built before we got scared, and so it got nothing. The task is not to fear the new thing more. It is to extend the discipline we built for the new thing back onto the old thing we stopped seeing.

The Coda, Returning to the Joke

So. The algorithm poops. I have now written two thousand-some words defending a bathroom joke, and I regret nothing, because the joke turned out to be the most honest description of the system I've found.

We spent a year and a series cataloging the appetite. We watched it eat virtue and language and love and democracy and reality and time, and finally turn on itself. What I missed, until the joke made me laugh, is that an appetite that large produces an output that large, and we have been swimming in the output the whole time. Calling it a feed. Calling it the discourse. Calling it the culture. When it is really just what's left after the nutrients were extracted and the meaning was thrown away.

The good news, and after a year of this series I am relieved to find some, is that we already invented the conscience this creature lacks. We built it for a smaller, newer animal and then politely declined to notice that a much larger one had been loose in the house for fifteen years. The constitution, the audits, the interpretability, the red-teams, the scaling policies: none of it is science fiction. It exists. It is funded. It is just pointed at the chatbot instead of the colossus.

We don't need to invent the conscience. We need to feed it to the right organism.

And we should probably do it soon, because the groundwater is starting to taste like the feed.
