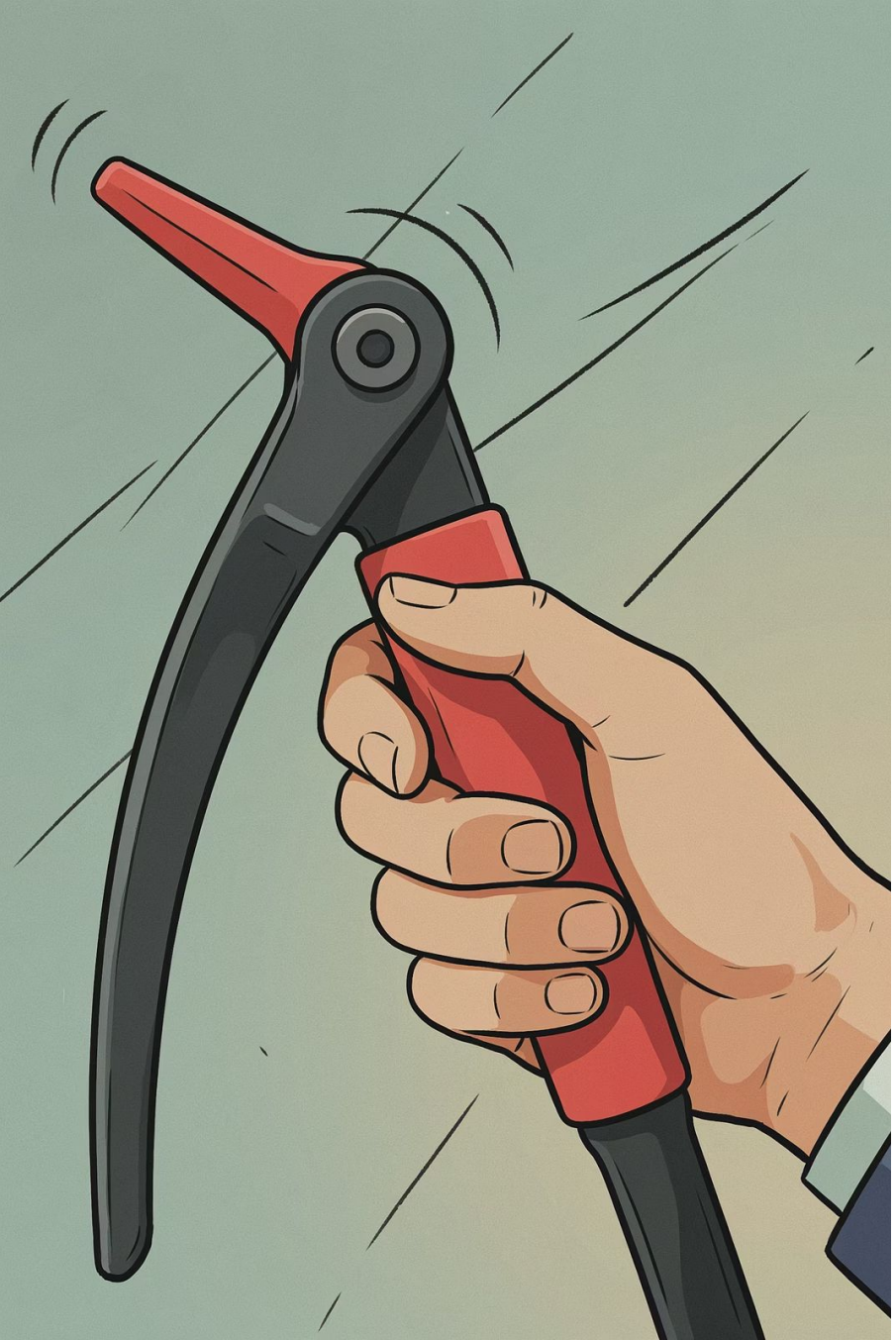


de-amplify: The Brake Integrity Standard

Regulate the loop, not the speech.

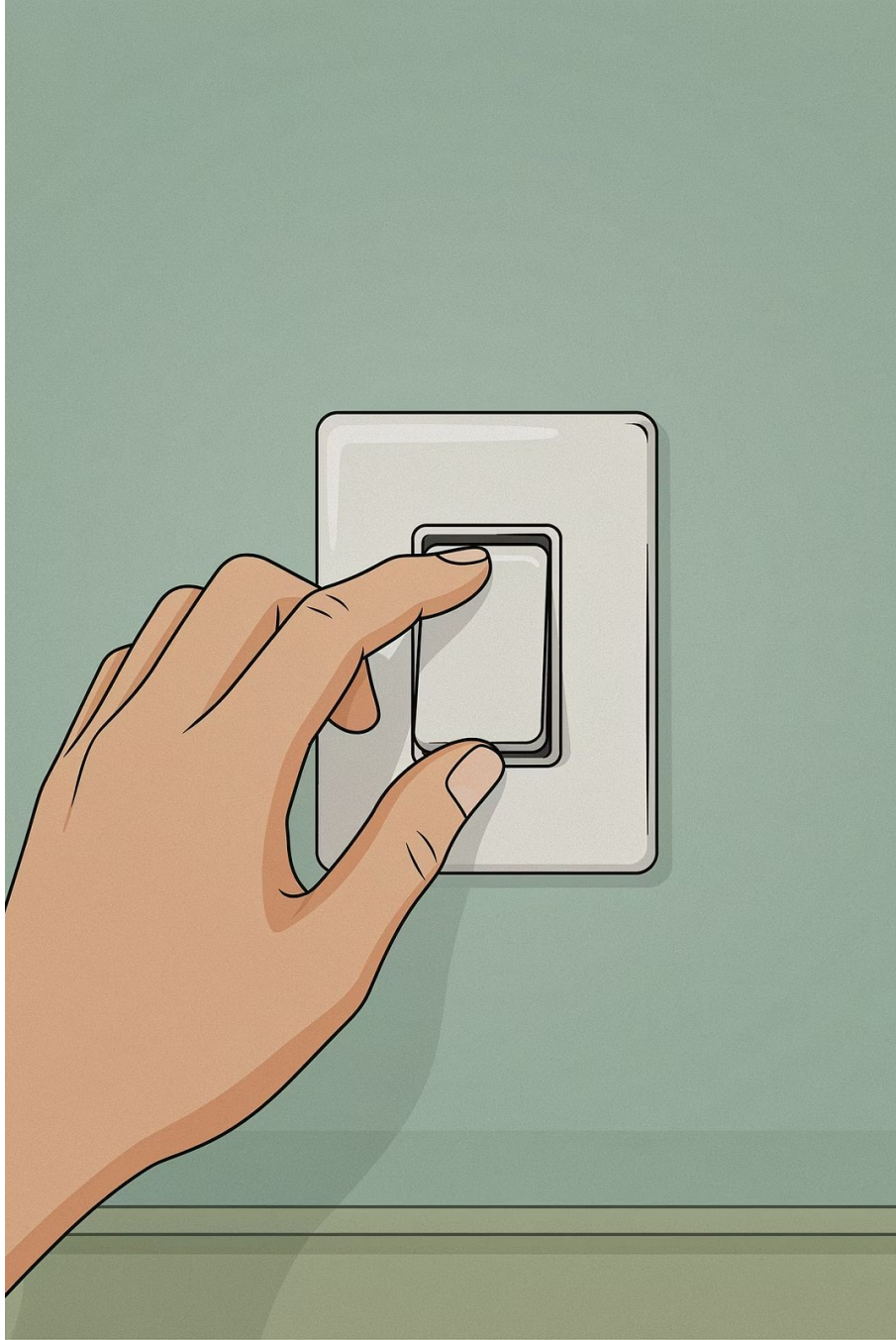
The same machine that hooks a child is the one amplifying division. Here is the part of it you can actually regulate — and it is not the content. It is the architecture of delivery and the controls over it.

The Problem, in One Sentence

A dead man's switch stops a machine when the human's hand comes off. Engagement feeds do the inverse: **the machine keeps running — or quietly restarts — after the hand comes off.**

> You choose "following only," close the app, reopen it, and you are back in the algorithmic feed. Your stated will was present in form, and absent in effect.

The control existed. It did not work. That gap — between a setting that appears to function and one that actually does — is the regulatory surface this standard targets.



The Bigger Machine Underneath

How the loop works

An engagement feed optimizes for one thing: whatever holds attention. High-arousal, outrage-driven content most reliably holds attention — so optimizing for engagement **structurally selects for the inflammatory.**

The machine has no opinion. It has your flinch, and it amplifies whatever you react to hardest.

The documented evidence

- Facebook's own internal research (2018) confirmed engagement ranking amplified divisive content
- The Haugen disclosures surfaced suppressed findings on harm
- Peer-reviewed literature since 2018 has corroborated both

The same loop that will not let a child stop is the loop that turns every political issue into a wedge. This is not only a youth-safety problem. **It is an information-environment problem.**

What Is Actually Broken

Not any single post. **The delivery loop.** A calm post and an inflammatory one ride the same architecture — and that architecture is the regulable surface.

Infinite Scroll

Removes natural stopping points, keeping users in the feed indefinitely.

Autoplay

Advances to the next piece of content without requiring any user decision.

Variable-Reward Notifications

Unpredictable pings that exploit the same behavioral loop as a slot machine.

Engagement-Optimized Ranking

Ranks content to maximize time-on-site and reaction, not user satisfaction or intent.

The more speech-defensible regulatory surface is **the architecture of delivery and the controls over it** — not the expression being delivered.

The Trap That Has Stalled Every Prior Effort

Swamp One: The Addiction Debate

"Is social-media addiction a real clinical condition?" leads to endless expert-witness battles over a contested diagnosis — a fight no regulator has won cleanly.

Swamp Two: The Content-Moderation War

"Which posts should come down?" puts the First Amendment directly in the path of every enforcement action. Courts have consistently made this the hardest terrain.

The reframe moves to something measurable and already inside consumer-protection law: **When a platform tells a user a control does something, does the control do it — and does it last?**

This Is Already in the Courts

Whether the platform's own brakes work is now **litigated evidence**, not a metaphor. Three active cases as of July 2026:

Case	Status
MDL 3047 (Federal) — ~2,893 consolidated suits	June 29 summary-judgment order sent states' claims to an August trial against Meta
K.G.M. v. Meta and Google (California)	$6 million jury verdict for negligent design; currently on appeal
New Mexico v. Meta	$375 million verdict; judge weighing a ~$953 million order that would change the product

In the federal order, the court found Meta's own documents could support the theory that its time-limit tools were a **"public relations stunt"** — the legal system's language for a brake that does not work.

Government Has Already Named the Design Surface

California SB 976

Restricts addictive feeds for minors. The addictive-feed provision and default-private-mode are **live, un-enjoined law** after the Ninth Circuit ruled in September 2025.

New York SAFE for Kids Act

Mirrors the California approach. Currently awaiting rulemaking to define enforcement specifics.

EU Digital Services Act

Mandates a non-profiled feed option. Preliminary findings issued against TikTok and Meta in 2026.

What none of them define: whether a user's decision to stop or redirect the feed **actually takes effect and persists.** That is the gap this standard fills.

The Standard: Brake Integrity

> When a service offers a control to stop, limit, reset, or redirect a personalization or engagement function, **the control must actually work.**

Layer I — Universal

Every account, any age, gets controls that work: turn personalized recommendations off, select a following-only feed, turn autoplay off, quiet engagement notifications. No age gate. No verification required.

Layer II — Youth Defaults

For minors, safer settings are **ON by default**, not merely available in a settings menu. The burden shifts from the child to the platform.

Making the base layer universal is deliberate: **the existence of a working brake never depends on age verification.**

The Compliance Spec: A Seven-Part Test

A control passes only if it satisfies all seven criteria. A cosmetic setting plus a press release fails on persistence and non-circumvention by design — and that is precisely the point.

Test	The control must…
Discoverable	Be findable without external instructions or support articles
Clear	State what will change — "show me less" is not "turn off recommendations"
Immediate	Change behavior soon after activation, not on some later refresh
Material	Change the feed perceptibly, not cosmetically
Persistent	Survive closing the app, another device, an update, and the passage of time
Scoped	Cover related surfaces — Reels, Shorts, Explore, notifications — not just one screen
Non-circumventing	Not nag, silently reset, or route the user back into the feed

The Wedge, Turned Into a Brake — Not a Label

Why labeling "divisive" content fails

- The party best positioned to build a divisiveness classifier **profits on the most inflammatory content** — Q1 2026 net income: $26.8 billion. It has every incentive to bias a silent classifier.
- A government-mandated label on lawful speech triggers **strict scrutiny** — the most legally exposed move available.

The defensible fix: user-held brakes

- **Transparency:** the feed must say why it served something — "you are seeing this because you engaged with similar posts."
- **A control over the signal:** the user can turn that recommendation signal off entirely.

That is recommendation literacy, not censorship. A brake on amplification, held by the user — not a labeler held by the state.

The EU's 2025 guidelines already recommend the "why you are seeing this" half. The contribution here is the **off-switch** — and making both legally testable.

Why This Survives the First Amendment

The core standard regulates the platform's **interaction with its own user**, not anyone's speech. Legal exposure runs from most to least defensible:

Tier	Approach	Legal Ground
1 — Safest	Control integrity: an offered control must actually work	*Lemmon v. Snap* lane — a design duty independent of content
2	Defaults and interface rules	Conduct-focused; SB 976's default-private provision survived Ninth Circuit review
3	Ranking-objective mandates	Meets *Moody v. NetChoice* head-on; more legally exposed
4 — Most Exposed	Content-characterizing labels	Strict scrutiny; SB 976's like-count ban was struck down on this ground

Brake integrity and the wedge off-switch both sit at **Tier 1.** Observable, defined criteria are written in on purpose: in *NetChoice v. Bonta* (March 2026), vague terms like "materially detrimental" and "well-being" were enjoined.

How You Enforce It

Observable Tests (Outside Access)

What it checks: Does the control activate, take effect, persist across sessions and devices, and stay scoped to the right surfaces?

Who runs it: FTC, state attorneys general, or an independent auditor — no platform access required for initial testing.

Mandated Internal Evidence (With Access)

What it checks: Control specification, version history, controlled-account testing results, and records of when a user preference was honored or overridden.

Who runs it: A digital regulator on the DSA model, with defined audit authority.

⚠ Anti-circumvention is non-optional: compliance must cover the **whole product**. If only the main feed is covered, the loop simply moves to notifications, streaks, and suggested accounts.

What Is Actually New Here

Existing law names design as the target. What has been missing are two specific tools:

1 A Testable "Does the Off-Switch Actually Work" Standard

SB 976, NY SAFE, and the DSA all name design as the problem. None wrote the **persistence and anti-circumvention test** that turns "a setting exists" into "the setting works, stays, and is checkable from outside the platform." This standard does.

2 The Wedge, Made Into a User-Held Brake

Transparency — why the feed served something — plus a control over the recommendation signal itself. The amplification answers to the person, not to a censor and not to the platform. This is the accountability mechanism that no current law requires.

Everything else in this framework — the universal base layer, youth defaults, the legal tiering — exists to make those two enforceable.

We Name Our Own Limits

A policy is stronger for saying what it has not solved. Four open questions, stated plainly:

Age Verification

A hard, privacy-fraught problem with no clean solution. **Mitigation:** Layer I is universal — the working brake never depends on age verification.

The Banishment Risk

A platform could stop serving minors rather than comply with youth defaults. The universal base layer is a partial hedge; the youth-defaults layer still creates a compliance incentive.

The Moody "Purely Reactive Algorithm" Question

Whether a purely reactive ranking algorithm can be regulated as conduct remains open — it has **not** been decided in this standard's favor. That exposure is real.

The Government-Mandated Divisiveness Label

This is the dangerous version of the wedge solution and is **deliberately NOT the ask.** This standard regulates the control, not the meaning of speech. That research stays in the appendix.

The Ask

01

Set the Universal Base Layer as a Floor

Any control a platform offers to stop, limit, or redirect an engagement function must meet the seven-part test — for every user, at every age.

02

Require Recommendation Transparency + User Control

Platforms must disclose why a post was served and provide a control to turn that recommendation signal off. This is the wedge brake.

03

Direct the FTC or a State AG to Publish the Audit Model

Observable tests and an audit protocol — not vague guidelines. Checkable from outside the platform, by public enforcers.

04

Write to the Vagueness Bar

Observable, defined criteria — never "well-being." Pilot the seven-part scorecard before thresholds go into statute. *NetChoice v. Bonta* is the warning.

The public is already here: **de-amplify.com** — a thirty-second self-test (find the brake, set it, close the app, reopen: did it hold?), constituent brake reports, and **#WheresTheBrake**.

The Standard, in One Line

> Existing law increasingly names design as the target. What is missing is a user-centered standard for whether the person's decision to stop or redirect the system actually takes effect and persists.

That standard is **brake integrity.** And the same brake, pointed at the recommendation signal, is how you de-amplify the wedge — without letting anyone censor it.

The Compliance Test

Seven observable criteria. Discoverable. Clear. Immediate. Material. Persistent. Scoped. Non-circumventing.

The Legal Architecture

Tier 1 conduct regulation. User-held controls. No content characterization. Survives the First Amendment.

The Public Entry Point

de-amplify.com — test your own brake, report the result, join the record.