Value Drift in Institutions

Joe Edelman

April 2, 2026

Each side is convinced the other is winning. The right noticed that universities, media, and cultural institutions no longer operate in service of excellence, or truth. Is it because the left captured them? The left sees courts, regulatory bodies, boardrooms, and governments no longer working towards fairness or the common good. Is it because the right captured them?

Both sides' values are losing, and each blames the other.

Here, I want to suggest somewhere else to put the blame: what's winning isn't anyone's values at all, but more like institutional purposelessness, busywork, nonsense. A kind of white noise.

This can co-exist with actual capture. For instance, maybe university professors are social justice lefties, who squeeze it into the curriculum. But professors don't set tuition, don't manage the endowment, don't run admissions. Those who do—do they foreground affordability? Access for the poor? Benefit to the public? Or do they optimize for *U.S. News & World Report* rankings? The left may have made some gains, but universities don't really serve left wing values, or right wing values either.

Same with the rightward turn in many nations' politics: Do governments now operate to increase right wing values like freedom, personal responsibility, and smaller government? Or do they optimize for re-election, media hits, or the interests of donors? Whatever captured the government, it's not exactly right wing values.

Clearly, some other force is in play. The numbers that go up seem unmoored from what people actually care about. Call it Goodhart's Law, the principal-agent problem, or value capture—but why does it happen?

My claim: institutions always drift from their values, but there are processes which, historically, have pulled them back. Those processes are failing right now, for reasons I hope to make clear.

How Institutions Drift

We should first be clear on why values drift starts. There are 3 main ways a gap opens between what an institution says it values, and what it actually does:

1. *There can be **inarticulacy** in the measurement system.* The institution's data collection and summarization systems may just not track the values they purport to serve.[1] A teacher can recognize education when it happens — say, when a student asks a question they couldn't have formulated a week ago. But that moment doesn't make it to the dashboard, often because metrics attempt to be behavioral, universal, and quantitative.[2] If the institution can't see what matters, it will optimize what it can see.[3]

1. *There can be **self-scoring** by those who set the metrics.* Managers and employees want to report success. Their bonuses, valuations, and headcount depend on it. If they can choose what success means, they'll pick metrics that are easy to improve, so that algorithmic changes, product features, or policy shifts can move the needle. Increasing time on site, or signups, may be easier than increasing real benefits.

2. *There can be **capture** by external actors.* An institution may start true to its mission, but then investors put their foot down, advertisers threaten to pull out, or some other pressure arises. This is legitimate when it adds a reasonable side-constraint (don't pollute, respect labor laws, submit to democratic oversight) even if these complicate the mission. But external actors often use their leverage to *replace* the mission's goals with their own. When the advertiser demands clicks over education, the institution is captured.

What Restores Institutions?

Inarticulacy, self-scoring, and capture happen all the time. But institutions have often corrected their course: the cronies were voted out, the metrics were revised, the

[1] In principal-agent theory, when a performance measure diverges from the true objective, optimizing for it distorts effort. Baker (1992) calls this *performance measure incongruence*.

[2] Behavioral metrics (clicks, completions, time on task) are *reasonless*: they record that something happened without capturing why and whether it mattered for the purposes of the person who did it, losing information. Organizations with products covering many use-cases and populations, will measure things common across them (votes, ratings, logins) rather than needs which differ, such as feeling heard, getting help, or learning something new. Finally, institutional mandates are often qualitative, while measurement happens via quantitative proxies.

[3] Holmstrom and Milgrom (1991) show that high-powered incentives on measurable tasks cause agents to neglect unmeasurable ones, hence the rationale for paying teachers flat salaries rather than incentivizing test scores.

institution returned to its former glory.

It comes down to consequences and people.

- **Consequences.** An institution with a bottom line that depends on real value (market share, popular approval, mission-critical outcomes) will get bitten by reality if infidelity goes far enough. Clients don't renew. Voters throw the bums out. Patients go to other hospitals. When the institution's survival depends on delivering real value, it has *skin in the game*.
- **People.** Or another thing that can correct drift is when people care. A hospital may be pushing to cut costs, but the doctor there still feels like the decision to treat or not is *hers*. She feels *moral weight*, a sense of *responsibility*, and she sees the patient, and knows whether the procedure helped. She has *direct visibility of consequences*.

In the ideal case, consequences and people drive successive clarification of values over time. The institution learns more about what *really* matters, re-aligns incentives, and re-engages with its mission on an ever-deeper level.

When Restoration Becomes Impossible

That restorative or deepening process doesn't seem to be happening as often as we'd like. In the corporate world alone: product teams optimize for engagement that doesn't track user value; sales teams chase quotas that ignore customer satisfaction. Parallels abound in academia (citations), journalism (clicks), education (certifications), medicine (throughput), and politics (poll numbers).

What's going wrong? I put it down to (1) bureaucratic insulation of decision-makers, which erodes moral weight and visibility of consequences; (2) the entrenchment of proxy metrics in contracts, which prevents revising inarticulacies, even when recognized; and (3) *value substitution loops*—self-reinforcing feedback between institutions and participants—which makes it impossible to reverse drift, even when everyone can see it.

Bureaucratic Insulation Historically, many institutions were smaller and more local. They had fewer layers of management, and often had more contact with those they served. In those smaller, local settings, you couldn't hide behind procedures (moral weight). Your actions were visible to you and your neighbors (visibility of consequences). And your reputation depended on delivering real value to the people around you (skin in the game).

But organizations scaled. Enter the *company man* — who hides behind procedures, is blind to real-world effects, and lives insulated from outcomes.[4] To the extent that modern institutions are staffed by "company men", they are more vulnerable to value drift, and less able to reverse it.

Contractual Lock-In Value drift becomes harder to reverse once proxy metrics get written into contracts (service-level agreements, procurement standards, funding requirements, regulatory frameworks, performance agreements). Those running a school district may know test scores aren't education, but if federal funding is tied to scores, or teacher evaluations are built around them, or parent expectations are shaped by score-based reporting, then renegotiating requires coordinated action across many parties, each already adapted to the proxy.

Value Substitution Loops At least with contractual lock-in, those involved can usually point to the gap between the proxy and the value, and the problem reduces to coordination — how to get all parties to move at once. The worst case is what I'll call a *value substitution loop*, where the strategic equilibrium moves such that the drifted state becomes the stable one.

It's easiest to show with a stylized example:

> A platform launches to help educators share what they know online. An educator joins to teach. The platform measures student engagement: watch time, comments, etc.
>
> Content that drives these metrics does well. The educator notices this and adapts, perhaps simplifying her takes, because to teach at all, she needs an audience. As thousands make this adjustment, the platform updates its model of what educators want. It sees that they are chasing engagement, and builds features to help them do that better.
>
> As the platform becomes better for chasing engagement, new entrants arrive with their eyes set on this. The role of "educator" has been redefined: being an influencer has become prerequisite to teaching. Even people with exactly the same values as the original educator try to become influencers first.

A proxy value (student engagement) has displaced the original one (education) through strategic adaptation by both participants and management. As participants adapt to the proxy, the institution adapts to serve them; as the institution adapts, participants adjust further. The result is a stable equilibrium, but one that serves nobody's values. It's not what the educators wanted. It's not what the students wanted. It's not even what

[4] A form of *moral hazard*, where insulation from consequences warps incentives.

the platform founders wanted. But no actor could revert to the original values without losing standing in the proxy-optimized system.

What distinguishes this from ordinary perverse incentives is that the institution must have some mechanism (algorithmic, bureaucratic, or market-based) that aggregates participant behavior and feeds it back into institutional design. Such a system has a tipping point once too many participants switch to the proxy-optimized strategy and the institution adapts to serve them.[5] It tips quicker when (a) dynamics for participants are competitive, (b) the institution adapts quickly (e.g., through algorithmic feedback), or (c) the proxy is far from the original value.[6] So irrevocable drift happens most in competitive, fast-moving, and hard-to-measure fields.

This of course happened in social media, but other examples abound: in academia, when researchers compete for tenure and grants, and optimize for citability via trendy topics and provocative framing, and then hiring committees see what gets cited and start favoring papers that generate buzz, this becomes what "good work" looks like. New PhD students are trained within this equilibrium.

Or in arts funding, for instance, Germany's experimental music scene: grant criteria highlight markers of "seriousness" like political framing (anti-colonialism, representation) or association with canonical avant-garde forms (free jazz, noise, electroacoustic composition). Artists present their work in these terms. Then funders narrow in even more. The drift is especially ironic as "experimental" gets redefined to mean preserving idioms which were transgressive fifty years ago.

The Human Cost

All this damages our society, of course. But there's also a personal toll: the researcher chasing citations; the teacher teaching to the test; the Instagram influencer living a lie. In uncharitable moments, we'd call them sellouts, but an injury is being done to them, and to all of us.

As Wolf Tivy put it:

> When I look at the things my friends were into before they destroyed themselves, this is what I see: false value sold to them by institutions and subcultures that have no structural reason to care about their real interests. But this applies to far more people than just the few that didn't make it. Almost everybody is trapped in some kind of propaganda complex, wasting their lives working for effectively nothing.

Life gets redirected away from what matters, toward what's easy to measure, easy to game, and easy to hide behind. If, as I mentioned at the beginning, both liberal and conservative values are losing, this is what's gained power: busywork, nonsense, "false value", and purposelessness.

This should be, at minimum, a reason for ceasefire. But I'd go further: If the real enemy is an institutional order that's lost contact with what anyone actually cares about, then let's rotate our turrets and aim squarely at that. Can we build institutions that stay connected to purpose? That measure what matters, keep decision-makers close to consequences, resist capture, and avoid value substitution loops?

I believe we can. I'll make that case in the next essay.

Thanks to Joel Lehman, Séb Krier, Ryan Lowe, Oliver Klingefjord, Toby Shorin, Ivan Vendrov, and Richard Ngo for comments, and Jamelle Watson-Daniels for generative discussions.

[5] More precisely: let $x \in [0,1]$ be the fraction of participants using proxy-optimized strategies and $\theta \in [0,1]$ be the degree to which the institution's design caters to proxy-optimization. Participants best-respond to θ; the institution best-responds to x. An agent adopts the proxy strategy when $\pi_P(x,\theta) > \pi_V(x,\theta)$, where π_P is the payoff to proxy-optimization and π_V to value-alignment. The institution updates $\theta = f(x)$ with $f' > 0$. Such a system has two locally stable equilibria — a "purposeful" one at low (x^*, θ^*) and a "drifted" one at high (x^*, θ^*) — separated by an unstable tipping point $\hat{x}$. What distinguishes this from standard principal-agent models is that in PA the principal designs the contract and the agent responds (Stackelberg); here the institution also adapts to agents. It is this co-adaptation that produces the trap. Cf. Bowles (1998) on endogenous preferences.

[6] Formally, $\hat{x}$ depends on three parameters: *competition intensity* c (how much participants must outperform each other to survive), *aggregation speed* α (how quickly the institution updates θ in response to x), and the *legibility gap* λ (the divergence between the proxy and the true value). The tipping point is decreasing in all three: $\frac{\partial \hat{x}}{\partial c} < 0$, $\frac{\partial \hat{x}}{\partial \alpha} < 0$, $\frac{\partial \hat{x}}{\partial \lambda} < 0$. So a discipline with 200 applicants per tenure line will tip into proxy-optimization at a lower fraction of defectors than one with 5; platforms with real-time algorithmic feedback should drift faster than institutions with slow feedback cycles (courts, churches), all else equal.