The Architecture of Feedback

Why equal error distributions produce unequal institutional drift

Abstract. Theories of organizational learning ask how an institution learns, imperfectly, from feedback it receives. Theories of organizational forgetting ask how it loses knowledge it once held. This article sits upstream of both. It asks which errors leave a trace at all, and holds that this filter is set by the architecture of the feedback, before any learning and before any forgetting. The central claim is conditional, and the condition is its strength: at equal initial error distributions, institutional policies diverge when their feedback architectures impose different binding forces on the local decision loop. Policies do not evolve from error distributions alone. They evolve through the architecture that turns some errors into binding feedback and leaves others mute. The reliability gate is the terrain on which the claim is demonstrated. The observability ratchet, treated in an earlier line of this work as the core mechanism, is here a corollary.

1. A principle, and a change of discipline it does not hide

The reliability gate was proposed, in the preceding installment, as a constructive device: a place where the compromise between safety and utility could be encoded by architecture rather than left to discretion. That proposal stands. This article does the opposite thing to it. It turns the device back on itself and watches it work, because a gate that functions perfectly can still produce a pathology, and the pathology is the subject.

The thesis first. A reliability policy does not drift according to how often it errs or how costly its errors are, taken as a distribution. It drifts according to how its architecture distributes feedback: which errors return a signal binding enough, and soon enough, to bear on the decision, and which return nothing at the moment that matters. Hold the distribution of errors equal across two gates, and they can still drift differently, because policy moves by the shape of the feedback its errors produce, not by their count.

The classical account is not blind to this, and it would be false to say so. A cost model can carry detection costs, delays, discovery probabilities, deferred costs, informational asymmetries. What it does not generally do is treat the asymmetry of feedback as an autonomous architectural mechanism of drift. It files the asymmetry as an exogenous parameter, a number handed in from outside, rather than as a structural property of the gate that generates the drift from within. Fields rarely forget variables; they file them in an annex. The contribution is to move this one from the annex to the mechanism, and to tie it to the architecture that produces it.

This article changes discipline as it goes, from the governance of AI reliability to a proposition about how institutions move their policies under asymmetric feedback, and

Jérôme Vétillard – VP R&D + Engineering / Chief Product Officer TweenMe® by Qualees®

the change is deliberate. Three claims are held apart by status. The *principle*, that at equal error distribution feedback architecture governs policy divergence, is general and conjectural. The *law*, its falsifiable specialization to systems governed by a local decision loop, is demonstrated on the bounded terrain of AI reliability gates. The *extension* to other domains is named as conjecture. The theory is, at this stage, more mature than its evidence, which is a single reliability instance, a body of clinical data gathered for other purposes, and a set of adjacencies in the literature. That is a common condition and stated once. The experiment that would move the law from stated to tested is specified in Section 6, and not performed here.

2. Three primitives, and one boundary

A theory is stable when it rests on few objects and derives the rest. Earlier drafts multiplied concepts until the vocabulary outgrew the claim. The correction is to reduce, and to fix, before reducing, the one boundary the argument turns on. In this article, *feedback architecture* denotes the local sociotechnical mechanism that turns a decision and its consequences into a signal addressed to the policy loop. An audit or a regulation is inside the architecture when it is part of that local mechanism and outside it when it is not. The boundary can include or exclude a given device depending on the system studied, but it is fixed for the length of an analysis, so that *endogenous* and *exogenous* do not become movable to taste.

The argument rebuilds on three primitives. The first is the *policy*, Π, the set of institutional parameters that govern the gate's decisions, the object that drifts. The second is the *binding force* of a feedback signal, B, the constraint a signal exerts on the local decision loop. The third is *institutional memory*, M, the formal reference of record. From these the rest is derived and should never be treated as fundamental. *Required*, the proof burden a policy demands at a port, is a projection of Π. The *relative cost of transformation* is a function of Π, M, and B. *Loud* and *Silent* are two regimes of B. *Down-typing* is a transformation of Π. The *ratchet* is a dynamic property of Π under a certain structure of M. No fourth primitive is admitted.

B needs care, because it is the object most exposed to the charge of being a convenient notation. It cannot be built from a list of incommensurable parts, latency in time, attribution in information, recurrence in counts, cost in currency, with no rule of combination; those four are its *determinants*, not its coordinates, and B is monotone in the obvious directions. It also cannot be a property of the signal, the architecture, and the organization all at once, stated interchangeably, and the resolution is a distinction inside B rather than a fourth primitive. Call B_A the *architectural binding capacity*, the constraint the architecture alone could impose, present wherever the architecture is and pointing one way. Call B_R the *realized binding force*, what that capacity becomes in a given organization at a given moment, raised or lowered by a right of override or a lax authority.

B_A is endogenous to the architecture; B_R is relational, B_R = B(A, S, O, t). The architecture owns the capacity, not the realization, and the article attributes to the architecture only the capacity.

Defining B by its effect demands a handle that is not the effect it predicts. The candidate handle is the *cost of suppression*, and it must be normalized or it measures the wrong thing. Raw suppression cost confounds the force of the signal with the rigidity of the governance around it: a weakly binding alert can be expensive to silence in a bureaucratic shop, and a strongly binding one silenced in two clicks in a lax one. The usable quantity is the suppression cost divided by the generic cost of changing the policy in that organization, C_s* = C_s / C_change, which strips out the administrative surround and leaves the part attributable to the signal. Even normalized it remains a proxy for the realized constraint, not a direct reading of B_A, and it is used as one. The experiment of Section 6 is what would give B empirical content in spite of the impurity.

3. From binding force to the cost engine, by mediation

Why does a binding signal produce a degradation of Required rather than an improvement of it? The arrow is real but not direct. It is mediated, and the mediation is where the argument earns its credibility.

A loud error, a false BLOCK or a REPAIR that halts the clinician in real time, produces a binding signal on the loop. The signal produces pressure for correction. The pressure raises the *perceived cost of holding the line*, of keeping Required where it stands while the friction continues. That perceived cost enters a choice, and the choice is not binary; an earlier draft made it too binary. Between raising the bar and lowering it lies a family of *containment* responses: suspending the system, narrowing its domain, escalating the case to a human, falling back to manual handling, reducing the volume sent through the gate. Containment holds Required in place while limiting exposure. The ratchet appears not simply when degradation is cheaper than promotion, but when degradation is cheaper than promotion and cheaper than every containment response available. Where a cheap containment exists, the loud friction is relieved without touching Required. Where it does not, degradation is the least-cost path, and the drift is the gradient of cost, followed.

This is the point to hold. The asymmetry is not a failing of the staff. It is the shape of the incentive field the architecture builds. A perfectly rational actor degrades Required under this field, because silencing the loud error is cheap and rewarded at once by the fall of friction, while raising the bar against the silent error is expensive and rewarded by nothing the loop can feel.

4. Memory by non-inscription

Here is the center of the argument, and the point at which it separates most cleanly from what precedes it.

A local decision, taken once, is not yet a norm. A threshold lowered on a Tuesday is an exception, so long as it stays an exception on the record. It becomes a norm at a precise moment: when the degraded state is inscribed as the reference of record, when the class-three attestation replaces the class-one verification as the artifact against which future compliance is judged. Institutional memory performs this inscription, and is therefore the level at which a local choice hardens into a standing rule.

The stronger claim is about what memory does not inscribe. An institution does not record what happens to it. It records what leaves a trace. The loud error leaves one by construction: it blocked someone, generated an override, produced an artifact that enters the record. The silent error leaves nothing. A false ALLOW, passed at runtime against a blind spot, produces no interruption and no document, and does not enter memory because it never produced anything for memory to hold. The bias of institutional memory is therefore not decay and not amnesia. It is a *non-inscription*. The record skews toward the loud not because the silent was forgotten but because the silent was never written. The proposition this yields is the most likely, of everything here, to outlast the article: an organization cannot learn institutionally from an event that never enters its system of proof, not because it chose not to, and not because it forgot, but because there was nothing to learn from where learning is kept.

Hold this against two neighboring bodies of work, since the resemblance is close enough to be mistaken for identity. Organizational learning, after Levitt and March (1988), takes the received signal as given and studies its bad interpretation. Organizational forgetting, after de Holan and Phillips (2004), studies the loss of knowledge once acquired. Non-inscription is upstream of both: not a bad reading of an available signal, and not the loss of a signal once held, but the structural absence of the signal at the threshold of the record, an absence the architecture decides in advance through the binding force it assigns. A precision the concept requires: a trace can be documentary, behavioral, metric, legal, narrative, or implicit in a routine, and a general theory of memory would admit them all. This article does not need them all. On its bounded terrain, the regulated reliability system, M is the *formal reference of record*, the audit artifact, the recorded threshold, the attestation kept as the standard against which compliance is judged, and restricting M to that record keeps it from swelling into an omnivore. In the vocabulary of this institute's prior work, the residue is the domain of *reconstructibilité*: what cannot be reconstructed after the fact, because nothing was laid down to reconstruct from.

Memory also bears on reversibility, though not in the simple way a first pass suggests. Memory does not, by its existence, make return expensive; a well-built memory can make it cheap, through versioning, decision history, rollback, a retained record of the old threshold. What makes return expensive is a particular structure of memory: the degraded state becomes the default reference, the prior state loses its standing, restoration requires a fresh authorization, and the burden of proof between lowering and raising is not symmetric. The irreversibility is not produced by memory as such. It is produced by the *asymmetry of re-inscription and restoration* that a given memory encodes. Naming that asymmetry names the remedy, and the remedy is more than instrumenting the absences. A memory built against the ratchet would version its policies, retain the prior level rather than overwrite it, require a signed justification for each degradation, expire exceptions automatically rather than let them settle into the baseline, hold the burden of proof symmetric, and keep a rollback within reach. The aim is not to record more. It is to make degradations non-self-perpetuating, so that a lowered threshold decays back toward the standard unless it is actively renewed, rather than hardening into it by default.

5. Two feedback regimes

The dynamics can now be stated with the economy the primitives allow, and the word attractor should be set aside; it promises a state space, transition equations, and stability conditions the argument does not supply. Call them feedback regimes.

The loud regime is *correction-seeking*, and the word is chosen with care, because an earlier draft called it self-correcting and claimed too much. The signal creates a pressure for correction; it does not guarantee the correction succeeds. The system may be unmodifiable, the cause external, the users may circumvent without improving, the managers may ignore the overrides. What the loud regime guarantees is a signal that bears on the loop, not a repair. Even so, the pressure is real and points toward correction, which is precisely what the silent regime lacks.

The silent regime is not correction-seeking, and the exact claim must be careful, because an earlier draft overstated it. The absence of correction does not by itself make the silent error grow. It makes it persist. Three states stay apart, and they separate two different dynamics. *Silent persistence* is the error remaining at constant policy: invisible, stationary, a latent liability. *Silent normative accumulation* is the policy dynamic, successive down-typings, each lowering the bar the next decision starts from, moving Π itself. *Silent exposure amplification* is the volume dynamic, more cases sent through the degraded policy without any further change to Π. The first is inert. The ratchet lives in the second and, through it, in the third, and the two should not be conflated, because one is a drift of the rule and the other a drift of what the rule is applied to.

The migration from the loud regime toward the silent one is what an earlier title named the ratchet. It is costly to remain in the loud regime, since remaining means paying the friction again and again; it is cheap to move toward the silent one, which is quiet and unmonitored. The system does not fail to correct the silent error through negligence. It is pulled, by the shape of its feedback and the asymmetry of its record, away from the region it can correct and into the region it cannot see. The irreversibility of that migration is a hysteresis in the strict sense of path dependence, not the magnetic hysteresis of physics: the state after a sequence of down-typings is not a function of current pressures alone but of the record those down-typings laid down, and that record does not clear when the pressure does.

6. A prediction, its measure, and the program that would test it

A theory becomes interesting when it predicts something the incumbent cannot, sharply enough to be wrong. Take two reliability gates with identical initial performance, identical precision, identical recall. The classical reading calls them equivalent. The law predicts that if they differ in the binding force their architecture assigns to the two error classes, they will drift at different rates and to different extents, initial performance notwithstanding.

The gap being measured needs its own definition, since a move from class one to class two is not a move from class two to class three, and an unweighted count would treat them alike. Let the promotion gap at time t be a weighted sum over ports, $G(t) = \Sigma w_i d(R_{i\_target}, R_{i\_actual})$, where d is a domain-defined distance between proof-burden classes, not a subtraction. The classes are ordinal, weights alone do not supply the missing metric, and d must be built from the costs or requirements the domain attaches to each class, or replaced by an ordinal indicator such as the weighted count of degradation transitions. Drift velocity is $v = \Delta G / \Delta t$; drift distance to a horizon H is $D_H = G(H) - G(0)$. The formalization is illustrative, and its service is to fix what would be recorded.

This is better read as an experimental program than a single experiment, and three difficulties keep it from being clean. Sampling the ALLOW decisions changes more than binding force: it changes the audit cost, the human load, the availability of labels, the frequency of correction, possibly the model itself, so the intervention does not isolate the feedback architecture perfectly. The conditions want randomization or simulation, since operators who know they are watched adapt to being watched. And the unit of analysis must be fixed, gate or port or decision or period or team or risk type, since the effect can appear at one grain and vanish at another. None of this makes the prediction untestable. It makes the test a designed program, which is the honest description of what a claim about institutional dynamics requires.

The single illustration available in the interim should be framed for what it is and no more. A reliability instance for toxicological prediction, built on a molecular featurization that encodes atoms by a one-hot atomic number, carries a structural blind spot for metallic coordination complexes, whose behavior depends on oxidation states and ligand chemistry the featurization does not represent. A confident false ALLOW on such a compound is a silent error in the exact sense used here: accepted at runtime, binding force near zero, no artifact, no entry in the record. This shows the silent class is not a theoretical convenience; it has a concrete instance with a concrete cause. One instance proves existence, not generality, which is why the program above is owed rather than assumed.

7. What this displaces, and which force is isolated

The comparisons are best read as a shift of causal priority rather than the discovery of empty ground. These bodies of work are not blind to the selectivity of feedback. Vaughan (1996) sees the non-event that fails to alarm; organizational safety has built whole apparatus around what gets reported and ignored. What this framework does is make that selectivity the primary object rather than one factor among many, and tie it to the architecture of the gate and the transformation of the policy. With that framing, each distinction is a change of emphasis and mechanism, not a claim of absence. Levitt and March describe learning from a received signal; here the object is the condition under which the signal exists. De Holan and Phillips describe the loss of acquired knowledge; here it is knowledge never inscribed. Automation bias, after Parasuraman and Manzey (2010) and Mosier and colleagues (1998), sits in the operator's cognition; here the level is institutional policy. Goodhart's law, in Strathern's (1997) formulation, describes a measure corrupted by becoming a target; here it is the port that degrades under a field the measure does not capture.

The near-miss literature needs a more exact placement than an earlier draft gave it. A near miss usually does leave a trace; the organization simply weights it lightly, because no serious consequence followed. The problem there is not an absent signal but a weakly binding one, which places near misses not at the silent extreme but in the interior of a continuum the theory should state plainly: a signal can be absent, present but weakly binding, or strongly binding, and the drift is governed by where on that continuum an error's feedback falls. The silent error is the limiting case where binding force reaches zero; the near miss is the instructive middle.

The clinical literature on alert fatigue belongs here as instance rather than rival. Reported override rates for medication alerts range from roughly forty-six to ninety-six percent, and institutions respond by suppressing alerts and reconfiguring thresholds. Read through the classical lens this is noise reduction; read through the law it is the ratchet observed in the field under another name, the suppression of the loud signal being the mechanism by which the policy migrates toward the quiet regime. The data are strong as illustration and weak as proof, since they were gathered to study fatigue and not drift.

The isolation of binding force needs a better defense than the earlier one. To call B structural and the rest perturbations was too convenient: regulation can be as structural as architecture, a budget as durably inscribed as a port. The honest distinction is not structural against perturbation but *endogenous against exogenous to the feedback architecture*, on the boundary fixed in Section 2. B_A is endogenous, a property of how the gate turns errors into signals. Regulation, budget, culture, leadership are exogenous, and may explain more of the variance in a given case than B does. B is isolated not because it dominates but because it is the one force the architecture itself produces, and therefore the one an architect can change.

8. The immune configuration, the scope, and the limits

The mechanism has a cancellation condition, and naming it does more work than any defense of the mechanism could. Let promotion be signed by an independent party, ALLOW decisions sampled systematically, the point of adjustment moved off the runtime into a distinct audit cycle. Then the exogenous restoring forces meet the endogenous one, the cost of degradation rises toward the cost of promotion, the absences are instrumented and memory ceases to be biased, and the drift does not begin. This is the cancellation condition and a hypothesis for design, not a clean refutation. A single system that satisfies the configuration and drifts anyway proves little, since the drift may come from a captured auditor, a falsified record, an undersampled review, a budget shock, any of the exogenous forces the law admits. Refuting the law needs a controlled comparison in which those forces are held still and only the feedback architecture varies.

The domains named earlier as sites of extension deserve a sharper description, which strengthens the argument. Aviation's near-miss reporting, nuclear precursor analysis, the just-culture reporting of high-reliability sectors are not empty ground awaiting the theory; they are the mechanism already engineered against, each an apparatus for giving a binding signal to the error that would otherwise pass silent. The comparison must separate three quantities, or it proves nothing: the nominal device, the real rate of inscription it achieves, and the real binding force it carries into the decision. A near-miss system with a low true inscription rate is an immune configuration in name only. The falsifiable corollary survives the caveat and is sharper for it: sectors that achieve high real inscription and high real binding should drift more slowly than sectors whose instruments are nominal.

On scope, the position holds: the principle stated generally, the law demonstrated only where its quantities can be measured, the extension offered as conjecture. One temptation is firewalled. It is tempting to conclude that institutions converge not toward the truth but toward whatever binds hardest, and to state this as a law. It is a doctrinal formula, kept as one, useful as compression and disqualified as authority; the law is conditional and holds only where a local decision loop governs policy adjustment. The limits are entered on the page: the theory is more developed than its evidence, which rests on one instance and one repurposed body of clinical data; the central program is specified but not run; the cost ratio is argued but not quantified; the binding force is given a proxy, the normalized cost of suppression, that is itself impure and unvalidated.

9. Conclusion

The reliability gate was offered, first, as a way to encode a compromise. It encodes something else as well, without anyone deciding that it should. By making one error loud and the other silent, it sets the binding force of their signals, and through that force it sets which errors reach the record and which never do. The policy then moves toward the errors it cannot ignore and away from the errors it cannot see, and the formal record, holding only what left a trace, turns the movement into a standing rule through the one asymmetry that matters: it is cheaper to write the degraded state down than to write it back.

The mechanism that gave this article its first title is the smaller part of what it now claims. The larger part is conditional and, being conditional, testable: at equal initial error distributions, institutional policies diverge when their feedback architectures impose different binding forces on the local decision loop. An institution does not record what happens to it. It records what leaves a trace, and something decides in advance what is allowed to leave one. That deciding thing is the architecture of feedback. It is worth naming, because a drift located in the architecture can be engineered against, rather than addressed through culture alone. A safety culture can change what gets reported, attributed, and inscribed, and to that extent it changes the architecture too. The claim is only that the architecture is where the drift is written, and therefore where it can be met.