Breaking the Monoculture: A Mandatory Non-Lineage Dissonance Seat for Self-Evaluating Multi-Agent Systems

Author: David Pineda **Affiliation:** OckhamLibra — Soviet Chess Framework **Date:** May 2026 **Keywords:** multi-agent systems, model monoculture, echo chamber, deliberation, ensemble diversity, Condorcet, Hong-Page, self-evaluation, LLM governance

---

Abstract

Multi-agent software systems built on a single model lineage share their failure modes: every agent's blind spots are correlated because every agent descends from the same training distribution. When such a system also *deliberates and votes* on its own improvements, voting does not correct the shared bias — it amplifies it with higher confidence. We report an empirical instance of this pathology and a disciplined remedy. In a production multi-agent framework where all six role-agents are instances of one model family, the formal deliberation mechanism (a four-voice "council") **never once fired across 98 tracked tasks**: ambiguity was silently absorbed in-context, the echo chamber operating without counterweight. We introduce the **asamblea retrospectiva**, a periodic self-evaluation in which each agent reflects on its own work anchored in deterministic telemetry, a strategist synthesises a triad of adjustment packages, and a **mandatory dissonance seat** — occupied by a *decorrelated, locally-hosted, non-lineage* model — objects to the triad before an evidence-anchored vote. The decisive design claim is that the seat's value is *independence, not capability*: in the first live run a 4-billion-parameter local model, strictly weaker than the orchestrating model, produced an objection that materially changed the outcome — downgrading a symptom-treating package and forcing a verification refinement that a same-lineage vote had been about to ratify. The system then self-corrected an instrumentation flaw that the assembly itself had surfaced, and the correction was verified against real data. We ground the mechanism in the Condorcet jury theorem's independence precondition and the Hong-Page "diversity trumps ability" result, and argue that any self-modifying agent collective must treat lineage decorrelation as a first-class, non-optional architectural element rather than a stronger model of the same kind.

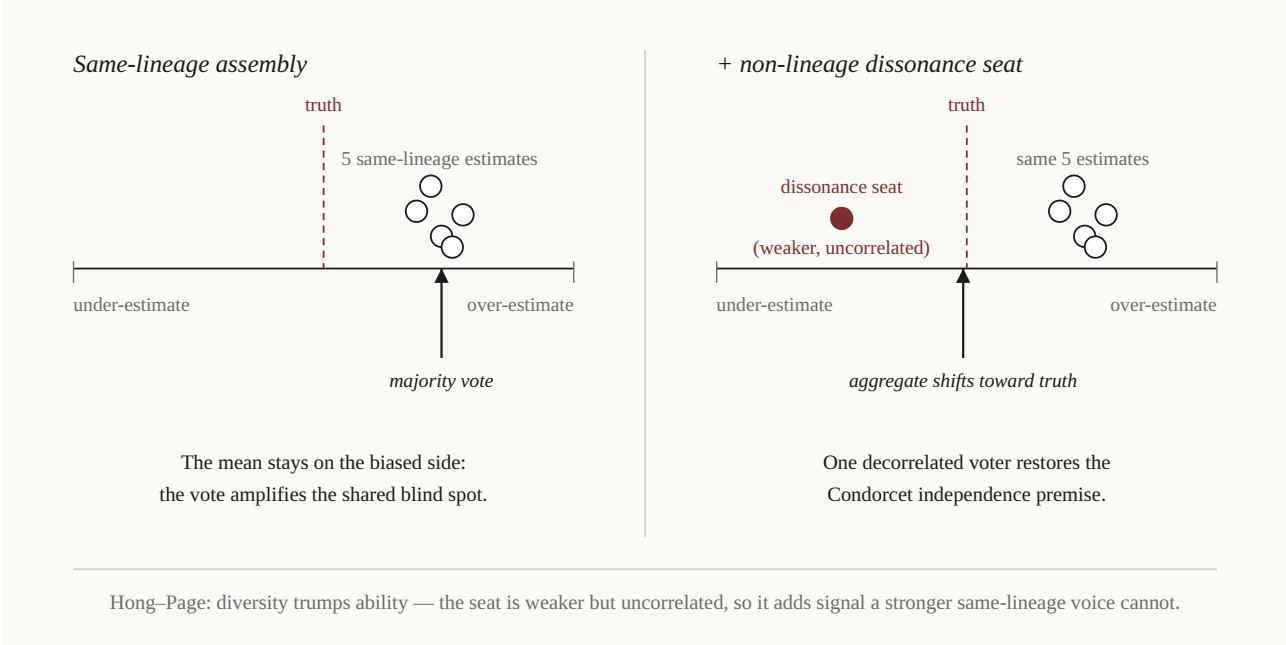
---

1. Problem: correlated blindness is not a capability deficit

A recurring assumption in agentic system design is that deliberation quality scales with model capability — that a stronger reviewer, a stronger planner, a stronger critic yields better collective decisions. This assumption fails for a specific and common class of systems: those in which every deliberating agent is an instance of the same model lineage.

The failure is not about how *good* each agent is; it is about how *correlated* their errors are. If reviewer, implementer, tester and orchestrator all descend from the same pre-training distribution and the same post-training, then what one cannot see, the others structurally cannot see either. Adding a stronger same-lineage agent reduces variance *in the wrong direction*: the collective becomes more articulate and more confident about the same blind spot. When the collective then **votes** on its own governance — which adjustments to

adopt, which process to change — the vote does not aggregate independent signal. It manufactures consensus. This is the multi-agent analogue of an epistemic monoculture.



Hong–Page: diversity trumps ability — the seat is weaker but uncorrelated, so it adds signal a stronger same-lineage voice cannot.

2. Empirical finding: the deliberation that never happened

The Soviet Dev Framework is a production multi-agent framework in which six role-agents (a chess-piece taxonomy: orchestrator, reviewer, backend, frontend, tester, release) coordinate through a persistent blackboard. It ships a formal deliberation skill — a four-voice *council* (Skeptic, Pragmatist, Critic, in-context synthesis) — explicitly intended for ambiguous, hard-to-reverse decisions, and whose own documentation names "premature consensus" as its primary anti-pattern.

Querying the blackboard's audit store across the full operational history yielded a stark result:

- **0 of 98 tasks** carried any `council:*` entry. The formal deliberation mechanism had **never fired**.
- The single traceable governance adjustment in the system's history had been produced by an *informal* deliberation, not the council.
- `0` keys for any dissonance / extended-critique mechanism the framework nominally provides.
- Meanwhile the reviewer agent had issued verdicts on 62 tasks (≈85% blocking historically) — decision pressure was high; convened dissent was absent.

The council had not failed loudly; it had simply never been reached, because its convocation criteria were strict and the orchestrator resolved ambiguity in-context. The echo chamber was not a hypothesised risk — it was the observed operating mode, invisible precisely because nothing was instrumented to make it visible. **The most important property of this finding is that it was telemetry-grounded, not introspective**: the system could not have discovered it by asking itself.

3. Mechanism: the asamblea retrospectiva

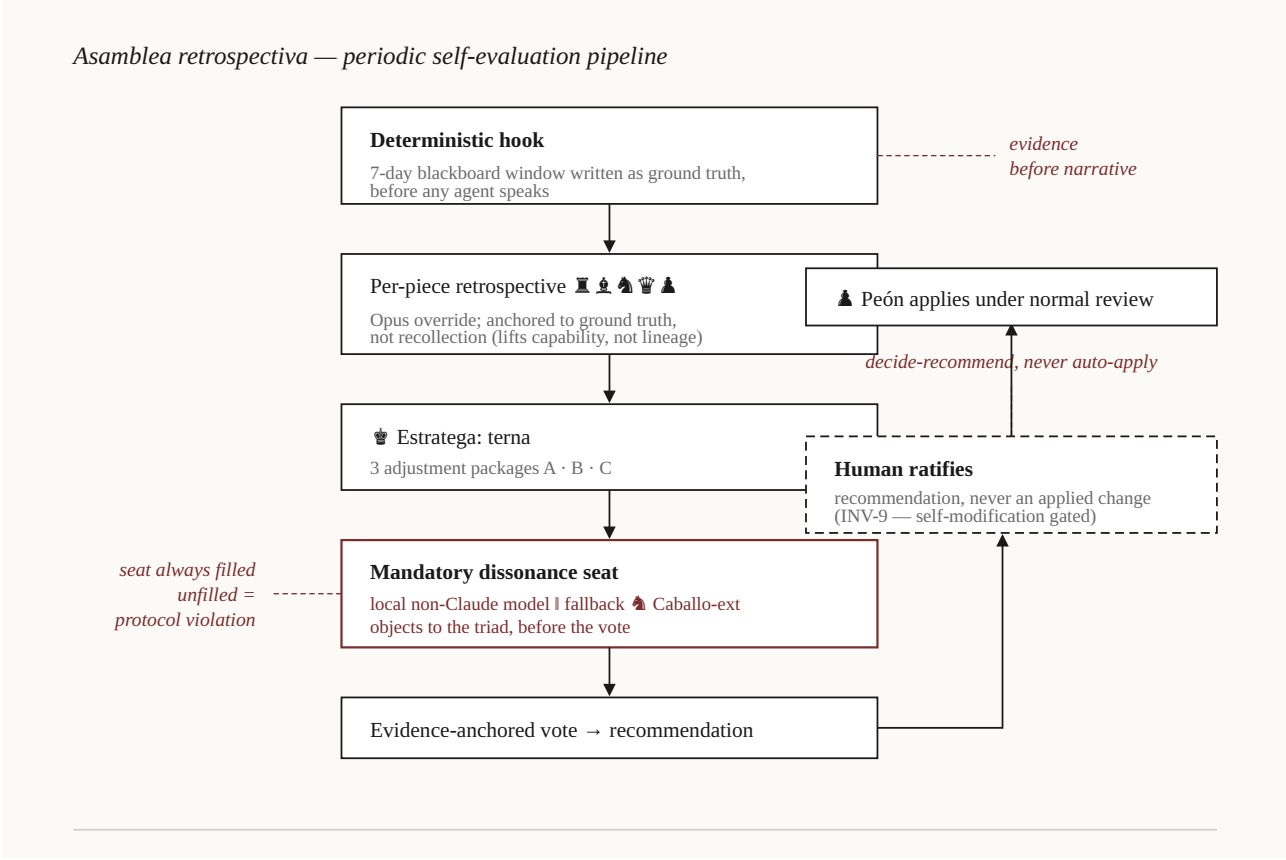
We introduce a periodic *retrospective assembly* — distinct from per-decision council — with four non-negotiable properties.

3.1 Evidence before narrative. A deterministic hook computes a fixed-window (7-day) retrospective directly from the blackboard (block/pass distribution, repeated-verdict tasks, reverts, per-agent operation volume) and writes it as ground truth *before* any agent speaks. Each agent's self-retrospective must be anchored to that record, not to its own recollection — closing the door on the favourable self-narration that same-lineage introspection invites.

3.2 Capability lift is acknowledged but explicitly insufficient. Agents reflect with a stronger model override than their normal operating tier. We state plainly that this mitigates *capability* and not *lineage*: a stronger Claude is still a Claude. This is why (3.3) is mandatory rather than optional.

3.3 A mandatory dissonance seat, occupied by a non-lineage model. Before the vote, a designated seat must object to the strategist's triad of adjustment packages. The seat is filled by a **decorrelated, locally-hosted, open-weights model of a different lineage** (in our deployment, a Qwen3 tag selected per-hardware; see §5). Its output is *advisory objection, never authority* — it is weaker than the orchestrator and must never decide or produce shipped artefacts. If no local model is available the seat degrades to an in-lineage extended-critique mode, but the seat is *always filled*: an unfilled dissonance seat is a protocol violation, not a no-op.

3.4 Decide-recommend, human-ratify. The assembly's vote produces a *recommendation*, never an applied change. A human ratifies; only then is each adjustment applied under its normal review discipline. Self-modification without a human gate is precisely the risk that disqualified a heavier autonomous-runtime alternative during design; the assembly is built so that the collective can propose changes to itself but cannot enact them.



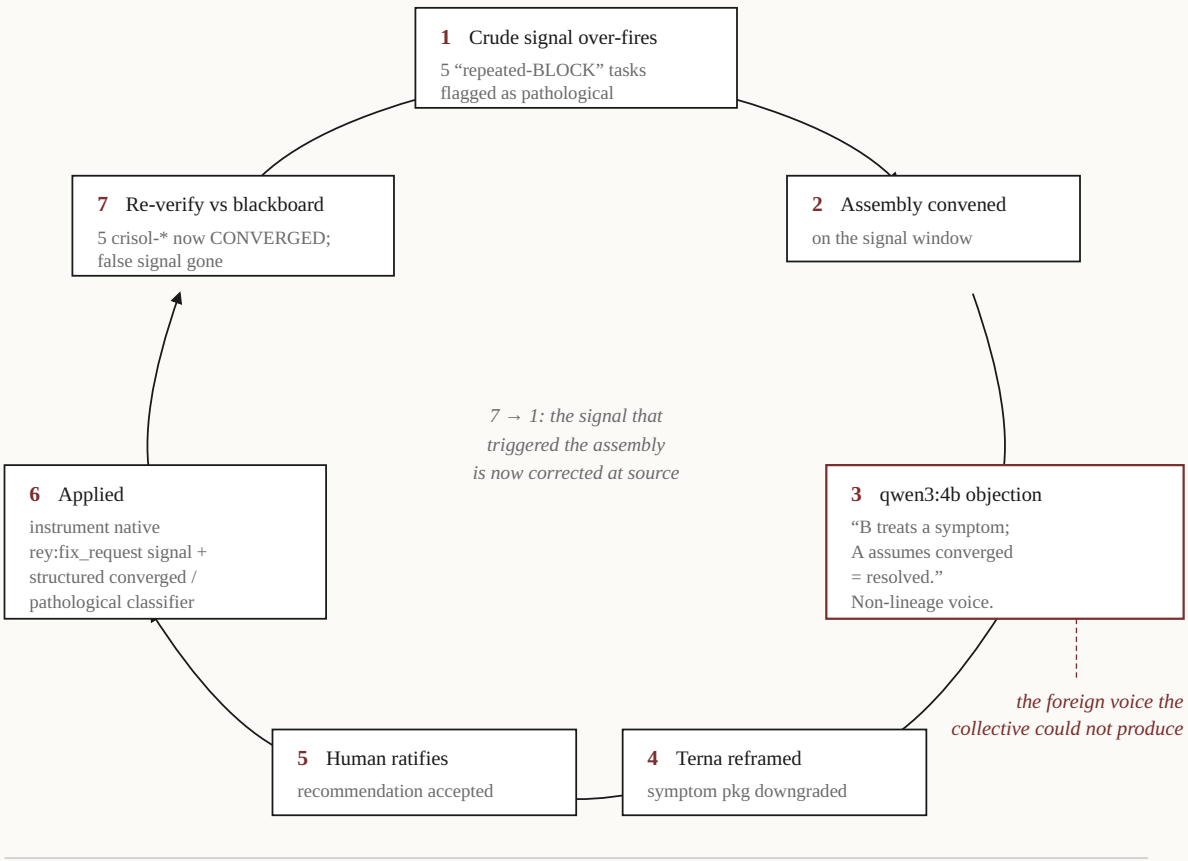
4. Result: a weaker model changed the decision

In the first live run the assembly convened on a real signal window. The strategist produced three internally-coherent adjustment packages. The dissonance seat — a 4-billion-parameter local model, strictly and obviously weaker than the orchestrating model — was given the triad and the ground truth and asked only to object.

Its objection was rough in prose but structurally decisive on three points: (i) one package *assumed* that a two-round "resolved" verdict was genuinely resolved rather than rubber-stamped; (ii) the packages treated agent overload and revert latency as independent when one plausibly caused the other; and (iii) one package was symptom-treatment — it mapped a symptom to a label without addressing why the underlying signal was invisible. The evidence-anchored vote, with that objection on the table, downgraded the symptom-treating package to a deferred dependency and reframed the winning package to include a verification step it had lacked. **A same-lineage vote had been on track to ratify the tidy original triad.**

The episode then closed on itself. The ground-truth signal that triggered the assembly was itself over-firing: five "repeated-verdict" tasks the heuristic had flagged as pathological had in fact all converged healthily. The winning package was precisely the fix for that crudeness. After human ratification the system instrumented the missing native signal and replaced the crude heuristic with a structured classifier; re-running against the real blackboard, the five tasks now classified as converged and the false signal disappeared. The assembly had used a non-lineage objection to detect the crudeness of its own instrumentation, proposed the correction, had it ratified, and verified the correction against real data — a complete self-improvement cycle whose integrity depended on the one voice that was not of its lineage.

A complete self-improvement cycle



5. Why a weaker model is the right occupant

The result is counter-intuitive only if one expects the seat to *out-reason* the collective. It does not. The relevant theory is about aggregation, not individual strength:

- **Condorcet jury theorem.** Majority aggregation improves on the individual only when voters are *independent*. Correlated voters make the vote amplify the shared bias. A same-lineage assembly violates the independence premise by construction; a non-lineage seat partially restores it.
- **Hong & Page, "diversity trumps ability."** A collective of diverse-but-weaker problem solvers can outperform a collective of strong-but-similar ones. The seat contributes *decorrelated error and decorrelated insight*, not superior judgement.

The operational consequence is sharp: do not staff the dissonance seat with a stronger model of the same kind, and do not let a weaker model's objection become authoritative. Its job is to surface what the correlated voices cannot, advisorily; the collective and the human still decide.

Locality matters independently. A self-hosted model adds zero data-exfiltration surface, no recurring cost and no rate limit; a hardware-adaptive selector picks the largest viable non-lineage tag for the host so the mechanism is portable across heterogeneous machines without committing a machine-specific choice to shared configuration.

6. Threats to validity

The evidence base is a single framework and a single first run; the quantitative claim (that the objection changed the outcome) is causal-by-inspection, not a controlled comparison. The dissonance model's prose quality is low and its objections require interpretation; treating its output as anything but advisory would degrade decisions. The convergence between the assembly's self-finding and its own fix is intellectually satisfying but should be read as one validated instance, not a general guarantee. Longitudinal evaluation — does the seat keep changing outcomes, and in a direction humans endorse — is the necessary next step and is deliberately gated behind a measurable success criterion before the mechanism is considered load-bearing.

7. Conclusion

A self-evaluating multi-agent system that shares one model lineage cannot audit its own monoculture by introspection or by a stronger model of the same kind; the corrective signal must be telemetry-grounded and the corrective voice must be lineage-decorrelated. We showed an instance where the formal deliberation mechanism had never once engaged across the full operational history, designed a retrospective assembly whose dissonance seat is mandatory and non-lineage, and observed a strictly weaker local model change a governance decision a same-lineage vote was about to rubber-stamp — then watched the system correct an instrumentation flaw the assembly itself surfaced, verified against real data. The contribution is not a model or an algorithm but an architectural stance: in any agent collective permitted to propose changes to itself, decorrelation is a first-class, non-optional element, and the cheapest, weakest, most clearly *foreign* voice in the room is the one that prevents the most expensive failure — the collective quietly agreeing with itself.

Companion: `dissonance-seat-paper-es.md` *(Spanish). Mechanism:* `docs/asambleas/README.md`*, skill* `asamblea-retrospectiva`*, hook* `daily-assembly.sh`*, selector* `scripts/asamblea/select-dissonance-model.sh`*. First assembly of record:* `docs/asambleas/2026-05-16.md`*. Genealogy of the rejected heavier alternative:* `docs/external_dependencies.md` *(*`NousResearch/hermes-agent`*, rejected;* `ollama-dissonance-seat`*, active).*