Romper la monocultura: una silla de disonancia obligatoria y de otro linaje para sistemas multi-agente que se autoevalúan

Autor: David Pineda **Afiliación:** OckhamLibra — Soviet Chess Framework **Fecha:** mayo 2026 **Palabras clave:** sistemas multi-agente, monocultura de modelo, cámara de eco, deliberación, diversidad de ensamble, Condorcet, Hong-Page, autoevaluación, gobernanza de LLM

Resumen

Los sistemas multi-agente de software construidos sobre un único linaje de modelo comparten sus modos de falla: los puntos ciegos de cada agente están correlacionados porque todos descienden de la misma distribución de entrenamiento. Cuando además ese sistema *delibera y vota* sobre sus propias mejoras, el voto no corrige el sesgo compartido — lo amplifica con más confianza. Reportamos una instancia empírica de esta patología y un remedio disciplinado. En un framework multi-agente en producción donde los seis agentes-rol son instancias de una misma familia de modelo, el mecanismo formal de deliberación (un "council" de cuatro voces) **no se activó ni una sola vez en 98 tareas registradas**: la ambigüedad se absorbió silenciosamente en contexto, con la cámara de eco operando sin contrapeso. Introducimos la **asamblea retrospectiva**, una autoevaluación periódica donde cada agente reflexiona sobre su propio trabajo anclado en telemetría determinista, un estratega sintetiza una terna de paquetes de ajuste, y una **silla de disonancia obligatoria** — ocupada por un modelo *decorrelacionado, local y de otro linaje* — objeta la terna antes de un voto anclado en evidencia. La tesis central de diseño es que el valor de la silla es *la independencia, no la capacidad*: en la primera corrida real un modelo local de 4 mil millones de parámetros, estrictamente más débil que el modelo orquestador, produjo una objeción que cambió materialmente el resultado — degradó un paquete que trataba síntomas y forzó un refinamiento de verificación que un voto del mismo linaje estaba a punto de ratificar. El sistema luego se autocorrigió una falla de instrumentación que la propia asamblea había sacado a la luz, y la corrección se verificó contra datos reales. Fundamentamos el mecanismo en la precondición de independencia del teorema del jurado de Condorcet y en el resultado "la diversidad supera a la capacidad" de Hong-Page, y sostenemos que todo colectivo de agentes que se automodifica debe tratar la decorrelación de linaje como un elemento arquitectónico de primera clase y no opcional, en vez de un modelo más fuerte del mismo tipo.

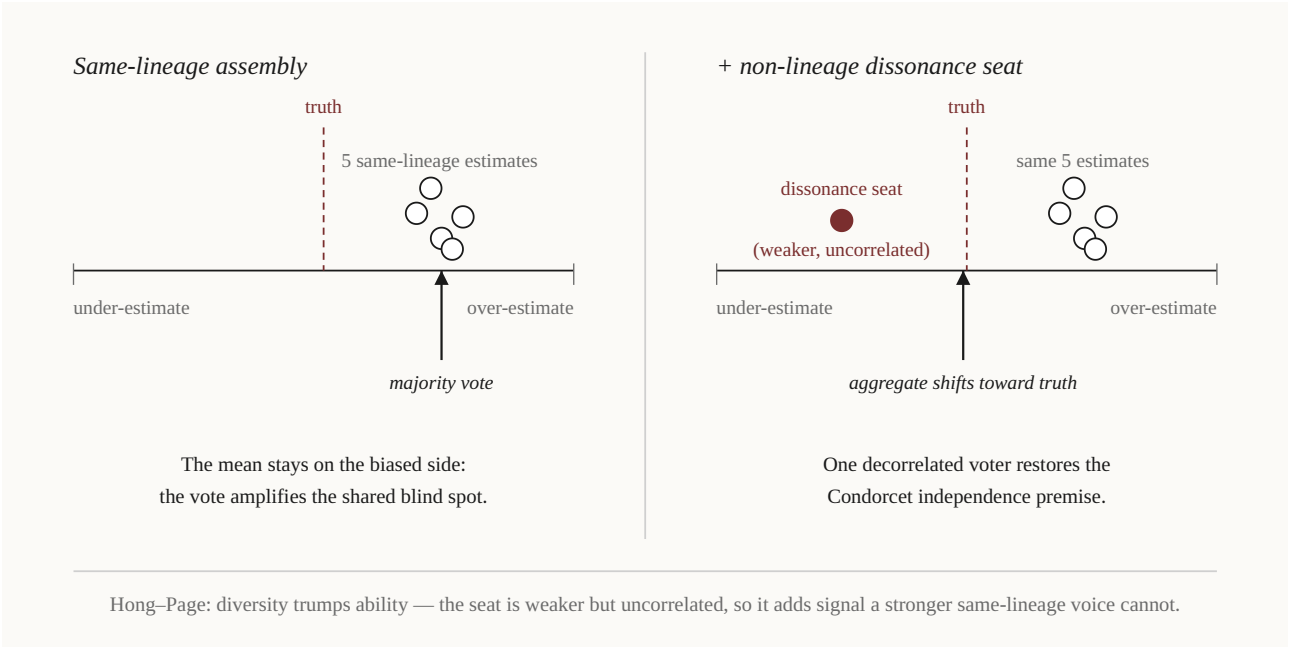
---

1. Problema: la ceguera correlacionada no es un déficit de capacidad

Un supuesto recurrente en el diseño de sistemas agénticos es que la calidad de la deliberación escala con la capacidad del modelo — que un revisor más fuerte, un planificador más fuerte, un crítico más fuerte producen mejores decisiones colectivas. Este supuesto falla para una clase específica y común de sistemas: aquellos donde cada agente que delibera es una instancia del mismo linaje de modelo.

La falla no es sobre qué tan *bueno* es cada agente; es sobre qué tan *correlacionados* están sus errores. Si revisor, implementador, tester y orquestador descienden del mismo preentrenamiento y el mismo

postentrenamiento, lo que uno no puede ver, los otros estructuralmente tampoco. Agregar un agente más fuerte del mismo linaje reduce la varianza *en la dirección equivocada*: el colectivo se vuelve más articulado y más seguro sobre el mismo punto ciego. Cuando luego el colectivo **vota** su propia gobernanza — qué ajustes adoptar, qué proceso cambiar — el voto no agrega señal independiente. Fabrica consenso. Es el análogo multi-agente de una monocultura epistémica.



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

2. Hallazgo empírico: la deliberación que nunca ocurrió

El Soviet Dev Framework es un framework multi-agente en producción donde seis agentes-rol (una taxonomía de piezas de ajedrez: orquestador, revisor, backend, frontend, tester, release) se coordinan mediante un blackboard persistente. Incluye una skill de deliberación formal — un *council* de cuatro voces (Escéptico, Pragmático, Crítico, síntesis en contexto) — pensada explícitamente para decisiones ambiguas y difíciles de revertir, y cuya propia documentación nombra al "consenso prematuro" como su anti-patrón primario.

Consultar el almacén de auditoría del blackboard sobre toda la historia operativa arrojó un resultado tajante:

- **0 de 98 tareas** tenían alguna entrada `council:*`. El mecanismo formal de deliberación **nunca se activó**.
- El único ajuste de gobernanza trazable en la historia del sistema había sido producido por una deliberación *informal*, no por el council.
- `0` claves para cualquier mecanismo de disonancia / crítica extendida que el framework nominalmente provee.
- Mientras tanto el agente revisor había emitido veredictos en 62 tareas (≈85% bloqueantes históricamente) — la presión de decisión era alta; el disenso convocado estaba ausente.

El council no había fallado ruidosamente; simplemente nunca fue alcanzado, porque sus criterios de convocatoria eran estrictos y el orquestador resolvía la ambigüedad en contexto. La cámara de eco no era un riesgo hipotético — era el modo de operación observado, invisible precisamente porque nada estaba

instrumentado para hacerlo visible. **La propiedad más importante de este hallazgo es que estuvo anclado en telemetría, no en introspección**: el sistema no podría haberlo descubierto preguntándose a sí mismo.

3. Mecanismo: la asamblea retrospectiva

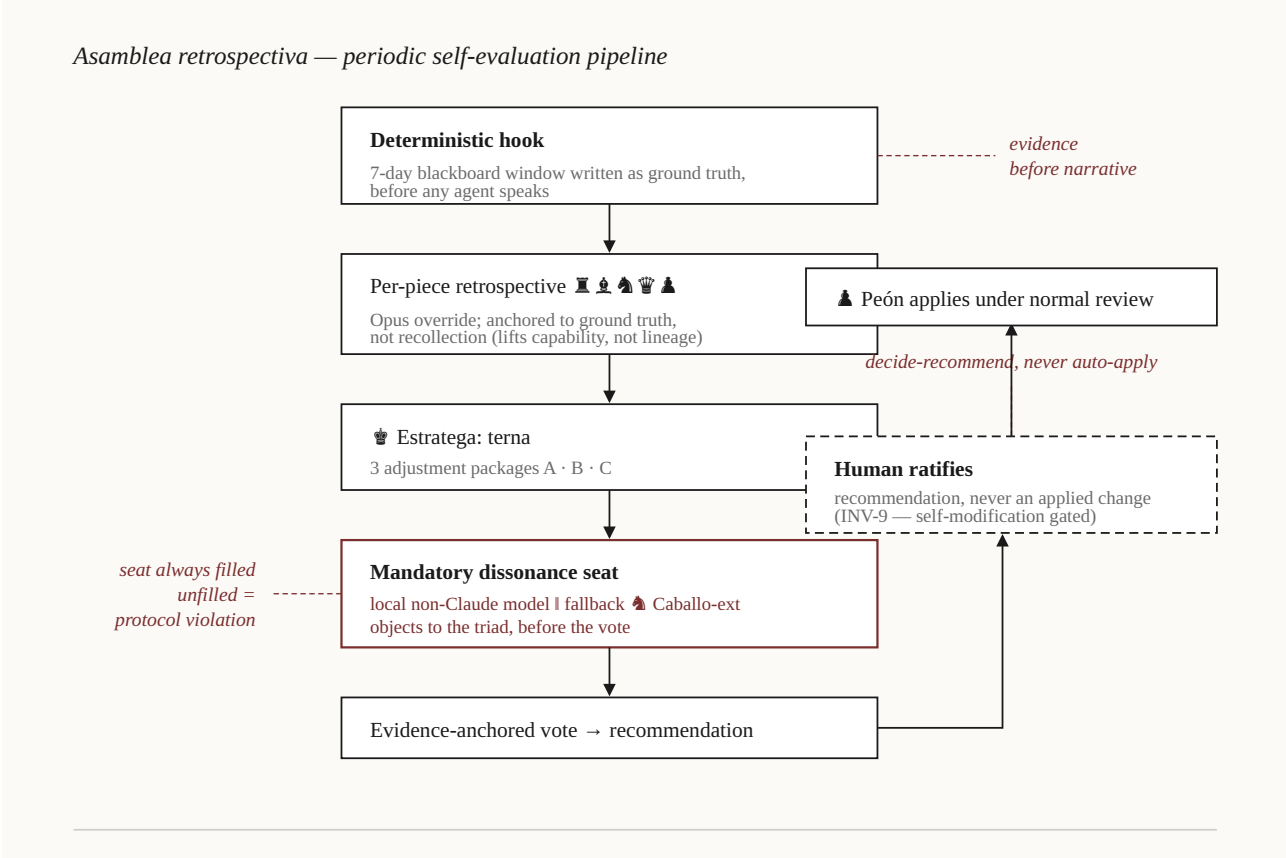
Introducimos una *asamblea retrospectiva* periódica — distinta del council por-decisión — con cuatro propiedades no negociables.

3.1 Evidencia antes que narrativa. Un hook determinista computa un retrospectivo de ventana fija (7 días) directamente del blackboard (distribución block/pass, tareas con veredictos repetidos, reverts, volumen de operaciones por agente) y lo escribe como ground truth *antes* de que cualquier agente hable. La autorretrospectiva de cada agente debe anclarse a ese registro, no a su recuerdo — cerrando la puerta a la autonarración favorable que invita la introspección de mismo linaje.

3.2 El salto de capacidad se reconoce pero se declara insuficiente. Los agentes reflexionan con un modelo de mayor calidad que su nivel operativo normal. Decimos sin rodeos que esto mitiga *capacidad* y no *linaje*: un Claude más fuerte sigue siendo un Claude. Por eso (3.3) es obligatorio y no opcional.

3.3 Una silla de disonancia obligatoria, ocupada por un modelo de otro linaje. Antes del voto, una silla designada debe objetar la terna de paquetes de ajuste del estratega. La silla se llena con un **modelo de pesos abiertos, alojado localmente, decorrelacionado y de otro linaje** (en nuestro despliegue, un tag Qwen3 seleccionado por hardware; ver §5). Su salida es *objeción advisory, nunca autoridad* — es más débil que el orquestador y nunca debe decidir ni producir artefactos que se envíen a producción. Si no hay modelo local, la silla degrada a un modo de crítica extendida del mismo linaje, pero la silla *siempre se llena*: una silla de disonancia vacía es una violación de protocolo, no un no-op.

3.4 Decide-recomienda, ratifica-humano. El voto de la asamblea produce una *recomendación*, nunca un cambio aplicado. Un humano ratifica; sólo entonces cada ajuste se aplica bajo su disciplina de revisión normal. La automodificación sin compuerta humana es precisamente el riesgo que descalificó una alternativa más pesada de runtime autónomo durante el diseño; la asamblea se construye para que el colectivo pueda proponer cambios sobre sí mismo pero no promulgarlos.



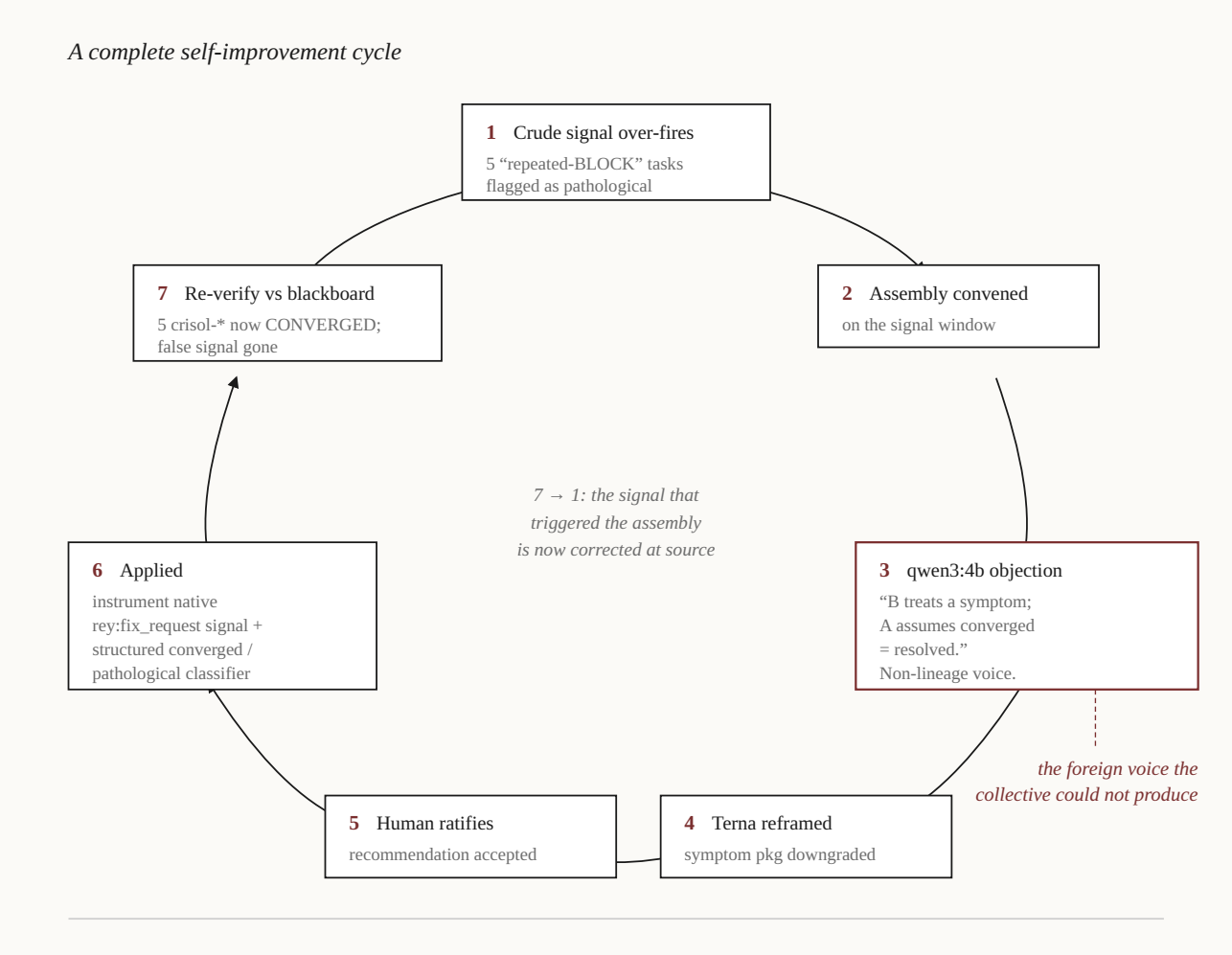
4. Resultado: un modelo más débil cambió la decisión

En la primera corrida real la asamblea se convocó sobre una ventana de señal real. El estratega produjo tres paquetes de ajuste internamente coherentes. La silla de disonancia — un modelo local de 4 mil millones de parámetros, estricta y obviamente más débil que el modelo orquestador — recibió la terna y el ground truth y se le pidió sólo objetar.

Su objeción fue tosca en prosa pero estructuralmente decisiva en tres puntos: (i) un paquete *asumía* que un veredicto "resuelto" de dos rondas estaba genuinamente resuelto y no sellado sin rigor; (ii) los paquetes trataban la sobrecarga de un agente y la latencia de reverts como independientes cuando una plausiblemente causaba la otra; y (iii) un paquete trataba síntomas — mapeaba un síntoma a una etiqueta sin atacar por qué la señal subyacente era invisible. El voto anclado en evidencia, con esa objeción sobre la mesa, degradó el paquete sintomático a dependencia diferida y reencuadró el paquete ganador para incluir un paso de verificación que no tenía. **Un voto del mismo linaje venía encaminado a ratificar la terna original prolija.**

El episodio luego se cerró sobre sí mismo. La señal de ground truth que disparó la asamblea estaba ella misma sobre-disparando: cinco tareas de "veredicto repetido" que la heurística había marcado como patológicas en realidad habían convergido sanamente. El paquete ganador era precisamente el arreglo de esa crudeza. Tras la ratificación humana el sistema instrumentó la señal nativa faltante y reemplazó la heurística cruda por un clasificador estructurado; re-corriendo contra el blackboard real, las cinco tareas ahora clasificaban como convergidas y la señal falsa desapareció. La asamblea había usado una objeción de otro linaje para detectar la crudeza de su propia instrumentación, propuso la corrección, la hizo ratificar, y verificó

la corrección contra datos reales — un ciclo completo de automejora cuya integridad dependió de la única voz que no era de su linaje.



5. Por qué un modelo más débil es el ocupante correcto

El resultado es contraintuitivo sólo si uno espera que la silla *razone mejor* que el colectivo. No lo hace. La teoría relevante es sobre agregación, no sobre fuerza individual:

- **Teorema del jurado de Condorcet.** La agregación por mayoría mejora al individuo sólo cuando los votantes son *independientes*. Votantes correlacionados hacen que el voto amplifique el sesgo compartido. Una asamblea de mismo linaje viola la premisa de independencia por construcción; una silla de otro linaje la restaura parcialmente.
- **Hong y Page, "la diversidad supera a la capacidad".** Un colectivo de solucionadores diversos-pero-más-débiles puede superar a uno de fuertes-pero-similares. La silla aporta *error decorrelacionado e insight decorrelacionado*, no juicio superior.

La consecuencia operativa es nítida: no dotar la silla de disonancia con un modelo más fuerte del mismo tipo, y no dejar que la objeción de un modelo más débil se vuelva autoritativa. Su trabajo es sacar a la luz lo que las voces correlacionadas no pueden, de modo advisory; el colectivo y el humano siguen decidiendo.

La localidad importa de forma independiente. Un modelo autoalojado agrega cero superficie de exfiltración de datos, costo recurrente nulo y sin límite de tasa; un selector adaptativo al hardware elige el tag de otro linaje más grande que sea viable para el host, de modo que el mecanismo es portable entre máquinas heterogéneas sin fijar una elección específica de máquina en la configuración compartida.

6. Amenazas a la validez

La base de evidencia es un solo framework y una sola primera corrida; la afirmación cuantitativa (que la objeción cambió el resultado) es causal-por-inspección, no una comparación controlada. La calidad de prosa del modelo de disonancia es baja y sus objeciones requieren interpretación; tratar su salida como algo más que advisory degradaría las decisiones. La convergencia entre el autohallazgo de la asamblea y su propio arreglo es intelectualmente satisfactoria pero debe leerse como una instancia validada, no una garantía general. La evaluación longitudinal — ¿la silla sigue cambiando resultados, y en una dirección que los humanos respaldan? — es el siguiente paso necesario y está deliberadamente condicionada tras un criterio de éxito medible antes de considerar el mecanismo como portante.

7. Conclusión

Un sistema multi-agente que se autoevalúa y comparte un linaje de modelo no puede auditar su propia monocultura por introspección ni por un modelo más fuerte del mismo tipo; la señal correctiva debe estar anclada en telemetría y la voz correctiva debe estar decorrelacionada de linaje. Mostramos una instancia donde el mecanismo formal de deliberación nunca se había activado en toda la historia operativa, diseñamos una asamblea retrospectiva cuya silla de disonancia es obligatoria y de otro linaje, y observamos a un modelo local estrictamente más débil cambiar una decisión de gobernanza que un voto de mismo linaje estaba por sellar — y luego al sistema corregir una falla de instrumentación que la propia asamblea sacó a la luz, verificada contra datos reales. La contribución no es un modelo ni un algoritmo sino una postura arquitectónica: en todo colectivo de agentes al que se le permite proponer cambios sobre sí mismo, la decorrelación es un elemento de primera clase y no opcional, y la voz más barata, más débil y más claramente *ajena* en la sala es la que previene la falla más cara — el colectivo estando de acuerdo consigo mismo en silencio.

Companion: `dissonance-seat-paper-en.md` *(inglés). Mecanismo:* `docs/asambleas/README.md`*, skill* `asamblea-retrospectiva`*, hook* `daily-assembly.sh`*, selector* `scripts/asamblea/select-dissonance-model.sh`*. Primera asamblea de registro:* `docs/asambleas/2026-05-16.md`*. Genealogía de la alternativa pesada rechazada:* `docs/external_dependencies.md` *(*`NousResearch/hermes-agent`*, rejected;* `ollama-dissonance-seat`*, active).*