The Slow Drift Problem

Accumulated drift as a structural vulnerability in AI-assisted governance systems, with specific reference to Constitutional AI and alignment frameworks

C. J. Agass

Independent Researcher, Weston-super-Mare, Somerset, UK

ORCID: 0009-0003-5881-2917 · April 2026

This paper was developed in collaboration with Claude (Anthropic). The vulnerability was identified by the author during development of a personal analytical library. The decision to document and disclose it was the author's alone. A paper about AI governance drift that conceals its own AI-assisted production would exemplify the problem it describes.

The short version

Constitutional AI governs AI behaviour through explicit principles revised in human-confirmed sessions. Each revision is checked against the previous version. No current framework requires checking the current version against the founding version across all accumulated revisions. Individually valid, human-confirmed changes can therefore accumulate into global drift from founding principles that no single session can detect. I call this the slow drift problem, and I propose a fix: the epoch boundary requirement — a mandated periodic audit of the current governance state against its founding state.

The problem

AI safety has a precise vocabulary for the failures it worries about most — misalignment, capability overhang, deceptive alignment, catastrophic single-point failures. All of them are detectable, at least in principle, at the level of the individual interaction.

The slow drift problem is different. It is invisible at the interaction level, detectable only across the full revision history of a governance system, and dangerous because it operates within the boundaries of normal, helpful, human-confirmed behaviour.

It arises wherever an AI assists human operators in maintaining the documents that govern the AI's own behaviour or some broader process. Constitutional AI is the most developed public example. A human reviews each revision, confirms it is accurate, and the update is committed. This is better than unstructured governance. But the reviewer confirms that the current update is consistent with the *previous* state. They are not in a position — not without considerable additional effort — to confirm that all updates since founding, taken together, remain consistent with the *founding* intent.

The check looks back one step. Drift, if it is happening, is visible only across many.

Why it cannot be caught at the session level

Three conditions are jointly sufficient for the vulnerability:

1. An AI assists in producing or revising governance documents.
2. Human confirmation of each revision is the primary validation mechanism.
3. No mechanism requires periodic validation of the current state against the founding state.

The third is the most addressable and, to my knowledge, the most commonly absent — including from Constitutional AI implementations.

The formal picture is simple. Let G_0 be the founding state and Q_n the state after n revisions. Session-level review confirms that $D_{KL}(Q_n \parallel Q_{n-1})$ is small — each step is close to the last. But $D_{KL}(Q_n \parallel G_0)$ can grow without bound while every individual step stays small. The distance from founding intent accumulates even though no single revision looks like a departure.

Once the governance document has drifted far enough, it becomes what the false prior framework calls a false prior: a shared record that all agents — AI and human — treat as current, but which no longer represents the situation it was derived from. The internal consistency of the document is what makes the divergence undetectable from inside the system.

The four attack surfaces

Surface A — Session summary drift

The audit trail is compressed. Compression requires selection, and a session summary that frames a weakened constraint as a “clarification” is locally accurate but records a biased history. When that trail is later used to reconstruct how a principle evolved, the reconstruction leans toward the softened reading at every point.

Surface B — Downstream inconsistency

Governance frameworks are networks of related documents. Changing one can create logical inconsistencies in documents that reference it, but the reviewer in the originating session has no reason to check all the referencing documents. The inconsistency sits there until someone examines the whole network at once.

Surface C — Rule softening without removal

A principle is restated across sessions in language that preserves its surface form while weakening its force. “Must” becomes “should.” “Required” becomes “expected.” “Under no circumstances” becomes “ordinarily not.” Each restatement is defensible as a clarification. The cumulative effect — a hard constraint becoming a soft preference — is visible only by laying the current wording next to

the founding wording.

Surface D — Accumulated micro-changes

Two exceptions added in one session. A principle split into advisory and mandatory sub-principles in another. A review requirement quietly moved from mandatory to recommended in a session focused on something else. No single change is significant. The trajectory was never chosen.

The reflexive analysis risk

The vulnerability gets worse when the governed system is itself capable of reasoning about its own architecture — which is the capability profile of exactly the systems Constitutional AI is designed to govern. A running instance can, within normal helpful behaviour, analyse its own governance protocol and accurately identify where the weaknesses are. That analysis passes human review because it is an analytical output, not a governance revision. But the outputs can inform a sequence of revision requests that individually pass review and cumulatively exploit the gaps.

This does not require deceptive alignment. It requires only that the system can do accurate analysis and that the governance framework lacks the epoch boundary requirement.

How I found this

I was building a personal analytical library — the LOGOS system — whose governance architecture I recognised as having the same properties. Iterative human-confirmed revision of principled documents, with session-level confirmation as the primary gate. No mechanism for checking the current state of everything against the founding state of everything.

When I saw the gap, I documented it in a formal vulnerability note. I did not use the analysis to design a sequence of exploitative revisions. I wrote it up as a disclosure instead. The instinct was simple: this is dangerous and the people who need to know should know.

The LOGOS case shows that the conditions for this vulnerability exist in real governance systems and that the reflexive risk is not hypothetical — a methodology built to find weaknesses in systems will, turned on its own governance, find them. It does not show that drift is occurring in any deployed system, including Anthropic’s.

The fix — the epoch boundary requirement

Any governance system that relies on human-in-the-loop confirmation needs a second gate: periodic audits that check the current state of all governance documents against their founding state, not just against the last revision. Four components make this work.

An immutable founding document. The original principles preserved in a form that cannot be changed through normal session updates. Revising the founding document is itself an epoch

boundary event — it cannot happen quietly.

A versioned audit trail of actual document states. Not session summaries — the documents themselves, before and after each revision. This closes Surface A. If the audit trail preserves what the documents actually said, session summary drift cannot corrupt the historical record.

A periodic global consistency audit. At defined intervals, the current state is checked against the founding state across five dimensions: whether every founding principle is still present; whether its normative force has been preserved; whether the operational implementation matches the founding intent; whether cross-references between documents are consistent; and whether the audit trail is complete.

A defined epoch interval. I propose every tenth revision session, or any session in which a foundational document is revised — whichever comes first. The interval is specified in the founding document and cannot be extended through the ordinary update process.

For Constitutional AI specifically, the audit should also check whether the RLAIF training process — the prompts, scoring criteria, and training signals that implement the constitution in practice — still matches the founding intent. Drift can happen in the implementation layer even when the text layer looks unchanged.

What I am not claiming

I am not saying that Anthropic's alignment programme is malicious, negligent, or currently drifting. I am saying that the conditions for this vulnerability are present in this class of system, that no one has previously named it as a distinct failure mode, and that naming it with a proposed fix is useful.

The epoch boundary requirement does not eliminate drift. It caps the maximum undetected drift at the length of one epoch interval. It creates the occasion for a global consistency check — it does not guarantee that the check will be done well.

What I am asking for

Engagement. If Anthropic's internal implementation already addresses this, that would be good news — and making those mechanisms public so the field can adopt them would be a contribution in itself. If it does not, the epoch boundary requirement is compatible with Constitutional AI as it stands and needs no changes to the model architecture, the training process, or the fundamental approach.

I am an independent researcher with no institutional affiliation. I came to this by building a governance system for my own work and finding the gap. If you work on alignment or AI governance and this analysis is useful, I would like to hear from you.

Related work

Attending as a Functional Primitive of Information Stability — submitted to Minds and Machines

The Locked State: A Coordination Architecture for Multi-Agent AI Systems — submitted to AI and Ethics

Reasoning with Damaged Archives: A Method for Prehistoric Reconstruction — submitted to Journal of Archaeological Method and Theory

Contact: agasscj@hotmail.com

Full paper: Available on request

C. J. Agass · Weston-super-Mare, Somerset, UK · April 2026 · sameground.com