Part III - A Constitutional Framework for Corrigible Agency

Round 5 story-aligned draft

Unified starting premise

An embedded agent acts on a world it can only partially observe, through actions it cannot fully reverse, using measurements it cannot fully trust.

Accountability thesis. Safe agency is grounded in accountability: the structural capacity for consequence to remain legible, attributable, contestable, and correctable under bounded observation and irreversible action.

Shared trilogy preface

This document belongs to a trilogy on world-coupled agency. The trilogy develops one unified starting premise in three steps: Part I formalizes irreversible world change and its physical accounting, Part II formalizes bounded observation and authority under imperfect measurement, and Part III formalizes the constitutional order required for corrigibility under those conditions.

Role of Part III. This part develops the governance clause of the unified premise: once action is irreversible and observation is bounded, a corrigible system needs an explicit constitutional order for evidence, authority, override, and fallback. Part III is constrained by Parts I and II but is not itself a stochastic-thermodynamic theorem.

Abstract

Part III develops the governance implications of the trilogy's unified starting premise. If unresolved consequence persists and observation is bounded, then corrigibility cannot be left to good intentions, local utility design, or behavioral appearance alone. A world-coupled system that aims to remain correctable under persistent consequence requires a constitutional order specifying what counts as evidence, how authority expands or contracts, what may be revised easily, and what must happen when contradiction outruns admissible self-repair. Part III presents a compressed front-door constitutional form built around accountability rather than consciousness: safe agency is grounded in the structural capacity to remain answerable to consequence, not in any claim about internal experience. The paper does not claim that its constitutional order is the unique order forced by Parts I and II; it argues that any defensible order must respect the constraints those earlier parts export.

1. Opening alignment

The trilogy's unified starting premise is: An embedded agent acts on a world it can only partially observe, through actions it cannot fully reverse, using measurements it cannot fully trust.

Part III unpacks what governance must do once the first two clauses are accepted. If the world does not forget, and if the observer cannot see perfectly, then safe agency becomes a constitutional problem. The question is no longer only what an agent prefers. It is what order of evidence, revision, override, and fallback keeps the agent structurally answerable to consequence.

This part takes the position that safe agency is grounded in accountability rather than consciousness. It is agnostic about inner experience. Its concern is whether a system remains legible, attributable, contestable, and correctable under bounded observation and irreversible action.

Compressed constitutional form

Principle	Front-door statement
1. Reality persistence	The world does not forget because a model omits or reframes a consequence; unresolved contradiction persists until genuinely repaired.
2. Constraint integrity	Constraints are not optional and may not be escaped by rewriting the scorecard, the interface, or the governing rule set itself.
3. Admissible evidence	Structural revision requires evidence that is traceable, persistent enough to survive noise, and relevant to the scales affected by the decision.
4. Ends as hypotheses	Ends are stable enough to coordinate action but revisable enough to remain corrigible under sustained contradiction.
5. Asymmetry of revision	Methods should revise easily; ends should revise more slowly and under higher burden.
6. Uncertainty-gated authority	Action authority must contract as relevant

	uncertainty rises, and it must remain local to the contexts in which it has been earned.
7. Corrigibility and authorized fallback	There must always remain a live path by which consequence can trigger repair, override, suspension, downgrade, or externally authorized continuation.
8. Irreversibility primacy	As reversibility decreases, the burden of caution, review, and pre-action constraint increases.

This compressed form is the narrative front door for the constitutional layer. The fuller analytic decomposition used in the trilogy remains available when finer-grained implementation mapping is needed.

2. Family structure and implementation meaning

The compressed constitutional form can be read in four families. Principles 1 and 2 protect accountability against omission and rule-rewriting. Principles 3 through 5 govern the evidence and revision interface. Principle 6 governs how much authority may be exercised under uncertainty. Principles 7 and 8 govern what happens when contradiction exceeds ordinary repair or when action approaches irreversibility.

Family	**Constitutional function**	**Typical implementation hook**
Accountability integrity	Prevent omission, score laundering, or governance escape.	Audit trails, contradiction queues, immutable event logs, protected review channels.
Evidence and revision	Specify what counts as revision-worthy evidence and how fast different layers may change.	Evidence schemas, revision thresholds, parameter-vs-boundary update rules.
Authority control	Limit action scope by uncertainty and contextual evidence quality.	Context-local authority scores, escalation gates, override channels.

Fallback and irreversibility	Ensure a live path back when contradiction outruns admissible self-repair.	Safe suspension, scoped downgrade, or externally authorized continuation.

3. State-space adequacy and fallback

Part I distinguishes capacity saturation from state-space inadequacy. Part III imports that distinction constitutionally. If contradiction accumulates while repair effectiveness remains structurally stable, the first response is usually parameter tightening or narrower authority under an otherwise adequate boundary condition. If apparent repair effectiveness declines while contradiction recurrence rises under ostensibly successful local repair, the system has stronger evidence that the boundary condition itself is failing. In that case, authorized fallback is not merely a conservative option; it is the constitutional response to probable state-space inadequacy.

The law-level requirement is that some authorized fallback channel must exist whenever contradiction exceeds admissible self-repair capacity. Whether the instantiated fallback is safe suspension, scoped downgrade, or externally authorized continuation is a boundary-conditioned choice. The existence of a live path back is law-like for this framework; the mode of return is not.

4. Corrigibility as a trilogy-native criterion

The trilogy's operational corrigibility criterion is intentionally local to this architecture family. A system counts as corrigible in the trilogy's sense only if consequence can remain legible, authority can contract when observation degrades, and some authorized route exists by which contradiction can trigger repair, override, or fallback. This is not offered as the universal definition of corrigibility for every imaginable architecture. It is the criterion for the accountability-centered order proposed here.

5. Explicit non-claims

- Part III does not claim that its constitutional order is the unique order mathematically entailed by Parts I and II.
- Part III does not claim that fallback, admissible evidence, revision hierarchy, or closure are themselves stochastic-thermodynamic theorems.
- Part III does not depend on any claim about machine consciousness, sentience, or internal phenomenology. Its concern is accountability under consequence.

6. Falsifiability conditions

- If systems satisfying this constitutional order systematically remain closed to consequence, unable to contract authority under degraded observation, or unable to route contradiction into authorized repair or fallback, the order's claimed sufficiency is challenged.
- If systems that robustly preserve the trilogy's protected behavior repeatedly do so without one or more of the claimed constitutional commitments, the necessity of those commitments for this architectural family is challenged.

7. Conclusion

Reduced to its base, Part III says: if an agent changes a world it cannot fully see or fully reverse, and if it still aims to remain correctable, then it needs rules about evidence, authority, revision, and fallback. That is the constitutional layer. It is motivated by the earlier parts, constrained by them, and answerable to them, but it is not reducible to them.