Accountability Under Partial Observation and Irreversible Action

Integrated flagship draft built from the trilogy root

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.

Abstract

This paper presents a first-principles framework for world-coupled agency built from one 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. From that premise the paper develops three linked layers. First, it states an accounting framework for persistent consequence and shows that, when the operational ledger is adequately modeled as a finite-state Markov jump process, the framework inherits exact stochastic-thermodynamic machinery, including the Master Equation, entropy production, generalized semantic observable currents, thermodynamic uncertainty relations, thermodynamic speed limits, and percolation-based containment conditions. Second, it shows that bounded and fallible observation implies context-local authority, uncertainty-sensitive action scope, and explicit protection against representation gaming. Third, it argues that any world-coupled system aiming to remain correctable requires a constitutional order for evidence, revision, override, and fallback. The paper does not claim to solve semantic alignment. It provides a first-principles framework for governing consequence, uncertainty, authority, and fallback in systems whose observables are semantically mediated and potentially imperfect.

1. Unified premise and accountability thesis

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.

The central claim of this paper is that corrigibility is best understood as a structural property of accountable world-coupled agency rather than as a behavioral appearance or a claim about inner experience. If the world can retain unresolved consequences, if observation is bounded and representation-dependent, and if agents still need to act, then any workable theory of safe agency must coordinate physics, epistemics, and governance rather than treating any one of them as sufficient on its own.

What the Markov layer adds

Layer 1 claim	Accounting statement	Layer 2 quantitative upgrade
Repair is not free	A world-coupled system cannot make contradiction disappear at zero cost.	For suitable currents on a finite-state Markov jump process, precision is lower-bounded by entropy production: Var(J)/<J>^2 >= 2/ΔS.
Delay worsens repair burden	Later repair is often broader-scope repair because propagation enlarges the effective burden.	Repair-time compression is lower-bounded by stochastic speed-limit structure: tau >= L^2/(2*Sigma*A_tau), under the stated regime assumptions.
Consequence can cascade	Cross-domain spillover can grow from local contradiction into system-level failure.	On the domain graph, containment is thresholded: below p_c spillover clusters remain finite; above p_c they can percolate system-wide.

2. Physical substrate: persistent consequence and exact model-class structure

At the broadest level, world-coupled action requires honest consequence accounting. Once an operational ledger is declared, unresolved consequence can change only through source, propagation, or realized repair. That accounting perspective is already enough to support

several qualitative claims: consequence persists unless genuinely repaired, repair is not a free sink, delay can enlarge the burden of repair, and cross-domain spillover can create cascade risk.

When the operational ledger is adequately represented by a finite-state Markov jump process, those qualitative claims acquire exact model-class structure. The Master Equation provides the continuity law; stochastic entropy production measures irreversibility; semantically weighted observable currents make explicit where semantic misspecification enters; thermodynamic uncertainty bounds convert repair precision into a cost floor; thermodynamic speed limits convert repair delay into a lower-bounded time-cost relation; and percolation theory gives a graph-level containment threshold for spillover.

Load-bearing object	Role in the framework
Master Equation	Native continuity law for the exact finite-state Markov instantiation.
Entropy production	Irreversibility measure and cost basis for precision and speed bounds.
Semantic observable current	Disciplined channel through which semantic weighting enters later governance without altering the raw generator.
TUR / TSL / percolation	Quantitative upgrades for precision-cost, time-cost, and containment claims already visible at the accounting layer.

3. Epistemic substrate: bounded observation and authority

The second layer begins from a simple question: how much action authority may an embedded agent earn from partial and fallible observation? This paper answers with three ideas. First, authority must be local to contexts in which evidence is actually earned. Second, authority must contract as relevant uncertainty rises. Third, a system must not be allowed to improve its apparent state merely by manipulating the representation through which that state is measured.

Epistemic commitment	Why it matters
Context-local authority	Evidence does not transfer uniformly across heterogeneous regions of state space.
Monotonicity under uncertainty	Seeing less should not license acting more.

Representation integrity	A system may not game its own scorecard by relabeling or suppressing the measurement channel.
Strict gating by default	High-impact action should remain limited by the most dangerous unresolved uncertainty unless a defensible completeness argument supports relaxation.

This layer does not solve semantic alignment. It formalizes the conditions under which semantically mediated observables may earn or lose authority. That is a narrower claim than solving alignment, but it is more operational and more honest.

4. Constitutional response: evidence, revision, and fallback

Once persistent consequence and bounded observation are granted, corrigibility becomes a constitutional problem. A system that aims to remain correctable needs rules about what counts as evidence, how authority expands or contracts, what can be revised quickly, what requires higher burden, and what happens when contradiction exceeds admissible self-repair.

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.

The constitutional layer is motivated by the earlier layers but not mathematically reduced to them. What Parts I and II force are constraints. What Part III adds is an accountable order for responding to those constraints.

5. Operational test program and model interface

The framework is designed to be testable. A working model need not demonstrate every philosophical implication to be valuable. It must, however, show that the framework makes contact with operational behavior. The highest-value demonstration is one in which the ledger is instantiated on a small finite-state process, semantic weighting is explicit, contradiction can propagate, authority can contract, and authorized fallback can be triggered.

- Physics-facing tests: check whether the declared model class supports the stated current, entropy, speed, and percolation relations at the chosen resolution.
- Epistemic tests: check whether authority contracts under rising relevant uncertainty and whether representation changes can be detected rather than silently rewarded.
- Constitutional tests: check whether contradiction can route into revision, override, suspension, downgrade, or externally authorized continuation when self-repair capacity is exceeded.

6. Falsifiability and limits

This framework is intentionally vulnerable to counterexample. If delay does not weakly increase effective burden under the paper's stated Layer 1 assumptions, the accounting corollary must be revised. If observed current statistics repeatedly violate the claimed TUR or speed-limit structure while the Markov regime claim is retained, the model claim or the export claim must fail. If authority increases under degraded observation, the epistemic layer fails. If contradiction cannot trigger authorized repair or fallback, the constitutional layer fails.

The paper also states a major limitation openly: it does not solve semantic alignment. The semantic observable current formalizes how semantic misspecification enters the observables that later layers must govern; it does not abolish that misspecification.

7. Conclusion

Reduced to its base, this paper says: world change persists, observation is bounded, and safe agency therefore depends on accountable structures for consequence, authority, and fallback. The integrated argument is cross-disciplinary, but its core is simple. Physics explains why contradiction cannot be wished away. Epistemics explains why authority must remain local and uncertainty-sensitive. Constitutional governance explains how a system remains correctable once the first two facts are accepted.

Drafting note

This flagship draft is built from the trilogy root and prior paper content, but it remains a story-first integration draft. Detailed references, a complete working-model section, and final venue-specific formatting should be added in the next pass.