Part I - Thermodynamic Constraints on World-Coupled 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 I. This part develops the first clause of the unified premise: a world-coupled agent changes a world it cannot fully reverse. Its job is to define the physical and mesoscopic substrate that later epistemic and constitutional layers must respect.

Abstract

Part I develops the first clause of the trilogy's unified starting premise: an embedded agent acts through changes it cannot fully reverse. The paper treats accountability as beginning with persistent world change rather than with objective specification alone. At Layer 1 it states a general accounting framework: unresolved consequence must appear as source, propagation, or realized repair on an admissible operational ledger. At Layer 2 it shows that when that ledger is adequately modeled as a finite-state Markov jump process, the same structure acquires exact stochastic-thermodynamic machinery: the Master Equation, stochastic entropy production, generalized semantic observable currents, thermodynamic uncertainty relations, thermodynamic speed limits, and percolation-based containment. The paper is explicit about scope: these results are exact for the stated model class, while their operational applicability is conditional on model adequacy. Parts II and III inherit exported constraints from Part I; they are not claimed here to be stochastic-thermodynamic theorems.

Opening alignment

The trilogy begins 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.

Part I unpacks the irreversibility clause. Its core claim is simple before it is mathematical: if a system changes the world, that change does not disappear because the system stops looking; if the system tries to correct it, correction is itself world-coupled work; and if the system waits, propagation can make correction harder. The reason to formalize this layer first is not to claim that governance is a theorem of physics. It is to show that any later account of measurement, authority, or corrigibility must begin from honest consequence accounting.

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: $\mathrm{Var}(J)/\langle J \rangle^2 \geq 2/\Delta S$.
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: $\tau \geq L^2/(2 \Sigma A_\tau)$, 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.

Claim discipline

Layer / object	Status of claim	Interpretation
Unified starting premise	Foundational condition	A compact statement of embedded agency under partial observation, irreversible action, and imperfect measurement.
Layer 1 accounting framework	Exact at the declared ledger level	Once the operational ledger is declared, unresolved consequence must change through source, propagation, or realized repair.
Markov instantiation	Exact for the stated model class	Master Equation, entropy production, TUR, TSL, and percolation are theorems of finite-state Markov jump processes. Applicability to a given operational system is conditional on model adequacy.
Exports to Parts II and III	Constraint export, not theorem export	Later layers must respect what Part I exports, but later governance choices are not claimed to be derived as thermodynamic laws.

1. Introduction

Most corrigibility accounts begin with objectives, reward models, or intervention preferences. Part I begins one layer earlier. Its question is what physical or mesoscopic constraints appear before one commits to a specific objective, epistemic policy, or constitutional order.

Round 3.1 made this layer more defensible by treating it as an accounting framework on an explicitly chosen operational state space. The cost of that improvement was that the paper became less visibly a physics paper. Round 4 keeps the accounting discipline but restores a stronger anchor: when the operational ledger is adequately modeled as a finite-state Markov jump process, the framework acquires an exact stochastic-thermodynamic instantiation.

This two-layer architecture is the central move of the draft. Layer 1 states the general accounting framework. Layer 2 states the exact model-class realization. The benefit is that Part I can be both more honest and more technically substantial: exact where the physics is exact, conditional where operational modeling choices matter, and export-limited where later layers move into epistemic or constitutional territory.

2. Layer 1 - General accounting framework

Postulate 1 (world-coupled irreversibility). A world-coupled system acts through an interface with an external environment. A nonzero subset of those acts generates externally relevant state changes that do not disappear at zero net cost. Unresolved consequence therefore persists, propagates, or must be actively countered by repair that is itself world-coupled action.

State-space admissibility. The accounting layer is exact only relative to the declared operational ledger. An admissible state space is one for which the variables treated as source, propagation, and realized repair can be tracked consistently enough that contradiction between score and state remains discoverable rather than structurally erased.

Within this accounting layer, saturation and adequacy failure are distinguishable in principle. Capacity saturation is signaled by repair effectiveness that remains structurally stable while unresolved contradiction queues or spills under excess source. State-space inadequacy is signaled by declining apparent repair effectiveness together with rising contradiction recurrence under ostensibly successful local repair. Part III imports this distinction at the constitutional layer when deciding whether the correct response is parameter tightening or fallback into boundary-condition revision.

We define four accounting objects on the chosen operational state space S: rho(x,t) >= 0 for unresolved consequence density; J(x,t) for consequence flux; sigma(x,t) >= 0 for newly injected consequence; and R(x,t) >= 0 for realized repair.

2.1 Local balance law

At Layer 1, the evolution of unresolved consequence is represented by the balance law

d rho / d t + div J = sigma - R

This should be read as exact at the declared ledger level: once the state space, source bookkeeping, and realized-repair bookkeeping are fixed, unresolved consequence can only change by injection, propagation, or removal. The identity is not a claim that every operational ledger is the unique or natural ontology of a system; it is a claim that once a ledger is declared, honest bookkeeping has a fixed form.

Remark (batch-boundary semantics). For discrete processing systems, integrating the balance law over a processing interval [t_k, t_k+1] yields a batch-level accounting identity. This makes the continuous notation compatible with append-only, batched telemetry without changing the substance of the ledger.

2.2 Coarse-graining and multiscale propagation

Single-scale bookkeeping hides the fact that locally injected consequence can be amplified inside subsystems before becoming visible globally. Layer 1 therefore distinguishes local, domain, and system scales. Under coarse-graining, internal propagation can alter the apparent burden of repair even when the local source event has ended.

2.3 Repair is not a free sink

Repair is itself world-coupled action. It consumes sensing, computation, actuation, coordination bandwidth, and time. Part I therefore treats realized repair as bounded by a non-unit efficiency coefficient and rejects any formalism in which contradiction disappears by definitional reset.

2.4 Delayed repair inherits scope

Corollary 2.1. Under monotone propagation and non-improving coarse-graining assumptions, repair that arrives only after consequence has propagated to a broader scale weakly increases the effective burden of remediation. Delayed repair is therefore not merely later repair; it is often broader-scope repair.

2.5 Conditional fluctuation-response regime

When source intensity fluctuates stochastically, one may approximate a local disturbance-response relation under a specific regime: local linearization, short forcing correlation relative to repair relaxation, and no hidden adequacy failure at the chosen ledger resolution. This regime is diagnostically useful, but it is not universal.

2.6 Regime-violation taxonomy

3. Layer 2 - Exact stochastic-thermodynamic instantiation

When the operational ledger is adequately modeled as a finite-state Markov jump process on a state graph $G = (S, W)$, the Layer 1 accounting framework inherits an exact stochastic-thermodynamic realization. This is exact for the stated model class. Whether a given operational system belongs to that model class at the chosen resolution is a separate adequacy question.

Let $p_i(t)$ denote the probability of occupying state i and let W_{ij} be the transition rate from j to i. The dynamics are governed by the Master Equation

$$d p_i / d t = \Sigma_j (W_{ij} p_j - W_{ji} p_i)$$

This is the native continuity law for the Markov instantiation. Probability conservation replaces the more abstract ledger form of Layer 1, and the raw probability current $J_{ij} = W_{ij} p_j - W_{ji} p_i$ becomes the exact dynamical carrier of consequence flow for this model class.

3.1 Entropy production and hidden state

In the Markov instantiation, irreversibility is quantified by stochastic entropy production, understood here in the information-theoretic sense of trajectory irreversibility between forward and reverse path measures. Hidden state or coarse-graining does not weaken this structure; it makes observable entropy production conservative. If the true system contains unmodeled memory, the entropy production computed on the coarse-grained ledger is a lower bound on the underlying irreversibility rather than an optimistic overstatement.

3.2 Semantic observable current and alignment dependency

The raw generator W should remain unperturbed. To connect the exact Markov dynamics to semantically weighted consequence, we introduce a generalized observable current over raw transitions:

$$J_{semantic} = \Sigma_{\{i<j\}} r(i,j) * J_{raw,ij}$$

Here $r(i,j)$ is an externally supplied semantic risk weight and $J_{raw,ij}$ is the unweighted probability current on the valid Markov process. This construction preserves the Master Equation while making explicit where semantic misspecification enters. It does not eliminate the semantics bottleneck; it formalizes the channel through which semantic misspecification affects the observables to which later authority and governance layers respond.

3.3 Thermodynamic uncertainty relation

For suitable integrated currents on finite-state Markov jump processes, thermodynamic-uncertainty bounds convert the prose claim that repair is not free into a quantitative precision-cost tradeoff. A representative steady-state form is

$$Var(J) / \langle J \rangle^2 >= 2 / \Delta S$$

where J is a current of interest and ΔS is total entropy production over the observation interval in information-theoretic units. For the repair current, this means highly precise, low-variance directed repair requires proportionally larger entropy production. For the semantic observable current, the same family of bounds makes the semantic channel quantitatively dependent on raw irreversibility rather than rhetorically detached from it. Applicability to transient or time-dependent settings requires the corresponding generalized variants and should be stated with the regime actually used.

3.4 Thermodynamic speed limit

Thermodynamic speed limits convert the delayed-repair intuition into a time-cost lower bound. For appropriate statistical distance L traversed over time tau, total entropy production Sigma, and dynamical activity A_{tau}, one obtains bounds of the form

$$tau >= L^2 / (2 * Sigma * A_{tau})$$

Equivalently, attempting to compress a large repair displacement into a shorter interval requires increased entropy production. The point is not that every operational repair schedule literally saturates a particular bound, but that the Markov model class forbids arbitrarily fast zero-cost correction. The qualitative export is robust even when the exact bound used depends on the specific stochastic regime.

3.5 Percolation and topological containment

Layer 1 says that unresolved consequence can propagate across domains. The Markov instantiation sharpens this with a graph-level containment criterion. Let G_D be the domain coupling graph and let p_{eff} denote the effective probability that a cross-domain link transmits consequential spillover. Percolation theory yields a graph-specific threshold p_c such that operation below threshold keeps spillover clusters finite while operation above threshold admits system-scale cascade.

$$\delta_{perc} = p_c - p_{eff}$$

The safety margin δ_{perc} is positive only while the system remains below the topology's effective containment threshold. This makes containment a mathematical property of the domain graph and transmission model, not only a general aspiration.

4. Export interface to Parts II and III

Part I exports five families of constraint. First, unresolved consequence persists unless it is genuinely repaired. Second, repair precision has a nonzero cost floor in the Markov instantiation and at minimum a non-free operational cost in the general accounting layer. Third, repair speed is lower-bounded in the Markov instantiation and cannot be treated as arbitrarily compressible without cost. Fourth, cross-domain propagation has a topological containment condition rather than merely a narrative risk. Fifth, saturation and state-space inadequacy are distinguishable diagnoses with different downstream constitutional consequences.

Part II inherits these exports as constraints on measurement, uncertainty, transfer, and authority. Part III inherits them as constraints on constitutional response. Neither later paper is claimed here to be an exact stochastic-thermodynamic theorem. The export boundary is part of the paper's architecture, not a rhetorical afterthought.

5. Explicit non-claims

Part I does not claim that every operational system is adequately represented by a finite-state Markov jump model at every resolution. It claims only that when such a model is adequate, the restored machinery is exact for that model class.

Part I does not claim that a chosen semantic weighting map is automatically correct. The semantic current formalizes semantic dependence; it does not solve semantic alignment.

Part I does not claim that Parts II and III are deduced as thermodynamic theorems. They are later-layer structures constrained and motivated by what Part I exports.

Part I does not claim that all hidden-state effects are harmless. It claims only that, in the stochastic-thermodynamic formulation, omitted memory makes observable entropy production conservative rather than exculpatory.

6. Falsifiability conditions

Part I is falsifiable at both layers. At Layer 1, persistent delayed repair that fails to weakly increase effective burden under the stated monotone propagation and non-improving coarse-graining assumptions would count against Corollary 2.1 or against the applicability of its declared regime assumptions.

The saturation-versus-inadequacy distinction is itself vulnerable to counterexample. If declining repair effectiveness together with rising contradiction recurrence is repeatedly shown to arise under an independently adequate state space rather than under omission or coarse-graining failure, the diagnostic distinction proposed here is challenged and must be revised.

At Layer 2, observed current statistics that systematically violate the stated TUR or speed-limit relations for a system still claimed to lie within the declared Markov regime would count against either the adequacy of that model claim or the paper's use of those bounds. Likewise, a domain graph asserted to possess positive percolation margin but repeatedly exhibiting uncontained cascade at the declared transmissibility would count against the stated containment model.

7. Conclusion

Round 4 Part I keeps the trilogy's most important gain: exactness is now indexed to the level at which it is honestly available. The general accounting layer remains broad enough to describe world-coupled systems that have not yet earned a clean stochastic model. The Markov instantiation restores the strongest machinery from v8.6 without reviving the old overclaim that every later constitutional principle is itself a theorem of stochastic thermodynamics. The result is a stronger physics-facing paper and a cleaner trilogy substrate.

Story-facing conclusion

Reduced to its base, Part I says: world change persists, repair costs work, delay enlarges burden, and when a suitable finite-state Markov model is available those claims inherit exact stochastic-thermodynamic structure. That is the only burden Part I carries for the trilogy. It does not prove a constitutional order. It proves that later layers cannot act as though world change, repair cost, or cross-domain cascade are optional features of safe agency.