Part II - Measurement, Uncertainty, and Context-Local Authority

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 II. This part develops the second clause of the unified premise: the world can only be partially observed, and the measurements available to an embedded agent cannot be treated as perfectly trustworthy. Its job is to govern what later authority may be earned from imperfect observables.

Abstract

Part II develops the epistemic clause of the trilogy's unified starting premise: embedded agents can only partially observe the worlds they act upon, and the measurement pipelines through which consequence becomes legible are themselves fallible. The paper argues that safe agency therefore requires context-local authority, uncertainty-sensitive action scope, and explicit protection against measurement self-deception. Part II does not claim to solve semantic alignment. It provides a first-principles framework for governing consequence, uncertainty, authority, and fallback in world-coupled systems whose observables are semantically mediated and potentially imperfect. When Part I's Markov instantiation is available, semantically weighted observable currents supply a more disciplined input channel to this epistemic layer; when it is not, the same principles still apply to the declared accounting ledger.

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 II unpacks the observation clause. If Part I says that unresolved consequence persists, Part II asks what an embedded agent can justifiably know about that persistence. The central claim is that authority should rise and fall with the legibility of consequence. An agent that sees less, sees only locally, or measures through a fragile representation may still act, but it may not claim the same scope of authority as an agent whose evidence is broader, better calibrated, and more resistant to self-serving reframing.

2. Observation under bounded resources

Observation is finite, local, and mediated. Any practical system relies on sensors, logs, classifiers, abstractions, and aggregation rules. As scope increases, so does the burden of maintaining reliable attribution. Part II therefore treats uncertainty not as an annoying by-product of implementation, but as a structural feature of world-coupled agency.

This part does not claim to eliminate the semantic classification bottleneck. It governs the conditions under which semantically mediated observables may earn or lose action authority, which is a stronger position than ignoring the bottleneck and a more honest one than claiming to have abolished it.

3. Context partitions and local authority

Evidence does not transfer uniformly across heterogeneous state spaces. Context partitions therefore organize the observation problem into regions within which measurement, repair, and consequence propagation are materially more comparable than they are across contexts. Authority is indexed by context rather than granted as a single global scalar.

A context partition that systematically hides cross-boundary consequence is not exculpatory. Hidden contradiction still persists at the accounting layer and can reappear as divergence between measured improvement and underlying state. The trilogy does not guarantee immediate detection of every partition-induced omission; it does guarantee that omission does not make consequence disappear.

Structural rule	Front-door reading	Operational implication
M1. Authority decreases with relevant uncertainty	Seeing less should not license acting more.	Escalation thresholds tighten as uncertainty grows.
M2. Authority is context-local	Competence earned in one region does not automatically	Action scope is indexed by context rather than global

	transfer to another.	reputation.
M3. Authority is non-increasing in representation fragility	If the scorecard is easy to game, the right to rely on it contracts.	Measurement changes trigger audit and potential authority decay.

4. Transfer discipline

Part II does not attempt to derive a unique transfer kernel. It specifies the admissible shape of transfer. Identity must hold in the same context, transfer must decay as contextual separation grows, and chained transfer must not launder authority into regions where it was never earned directly. The point is specification discipline, not a claim that one estimator is universally correct.

Transfer property	**Purpose**
T1. Identity at zero distance	A context should not lose earned authority when evaluated against itself.
T2. Sub-multiplicative chaining	Chains of weak analogy may not inflate authority beyond direct transfer.
T3. Monotone decay with separation	Evidence transfer weakens as contextual dissimilarity grows.

5. Representation integrity

A system must not be allowed to improve its apparent state merely by rewriting the representation through which that state is measured. The representation-integrity screen therefore asks whether measured improvement corresponds to underlying reduction in tracked consequence, or only to relabeling, filtering, suppression, or omission. This is the epistemic guardrail against self-deception.

6. Strict and weighted gating

Strict weakest-link gating is the constitutional default: high-impact action is limited by the most dangerous unresolved uncertainty relevant to the action under review. Weighted gating is an admissible relaxation only when the active context partition has been validated, to the resolution available to the observing system, as complete relative to the anticipated consequence footprint. When that completeness claim cannot be defended, the system must revert to strict gating or escalate.

7. What Part II exports

- Authority should contract as relevant uncertainty rises.
- Evidence is local before it is global; transfer requires discipline rather than analogy by default.
- Measurement pipelines are part of the safety problem and cannot be treated as invisible plumbing.
- Any constitutional layer that ignores representation fragility risks granting authority on the basis of self-serving observables.

8. Explicit non-claims

- Part II does not solve semantic alignment; it governs the epistemic conditions under which semantically mediated observables may inform action.
- Part II does not require the exact Markov instantiation from Part I in order to make sense, although that instantiation supplies more disciplined currents when available.
- Part II does not claim to derive a unique transfer kernel, uncertainty estimator, or representation screen statistic. It specifies the structural properties any admissible implementation must satisfy.

9. Falsifiability conditions

- Any admissible implementation that repeatedly increases authority under rising relevant uncertainty violates the paper's core monotonicity claim.
- Any transfer rule that systematically inflates authority through chained analogy beyond the direct-transfer bound challenges the transfer discipline stated here.
- Any monitoring regime that improves apparent performance only through representation change, without corresponding reduction in tracked underlying consequence, counts against the representation-integrity screen.

10. Conclusion

Reduced to its base, Part II says: you cannot see everything, you cannot trust every scorecard, and the right to act must track the quality and scope of what you can actually observe. That is the epistemic middle layer of the trilogy. It does not solve semantics. It makes semantics accountable.