Systemic Epistemic Governance

Deriving a Proposition versus the Authority to Persist and Act on It: An Operational Semantics for Provenance-Gated Admission in Persistent Cognitive Systems

Stefan Ragland

Dominion Labs Research & Development

Dominion Labs

research@dmnlabs.org

August 5, 2025

Abstract

A persistent cognitive system must decide, continuously, which of its own inferences may change its long-term state. We argue for a single principle: *a system should distinguish the ability to derive a proposition from the authority to persist and act upon it, and that distinction can be enforced as a system-wide state-transition invariant*. We call the resulting discipline *systemic epistemic governance* (SEG). SEG partitions persistent state into an *authoritative tier*—the store the reasoner treats as authorized and the rules permitted to act—writable only through provenance admission or *independent* validation, and a *soft tier* of beliefs, derivations, and generalizations that materializes automatically and is, by construction, non-authoritative and volatile. We give a small-step operational semantics for SEG and prove four *structural* guarantees—authoritative provenance-rootedness, non-circular promotion, execution safety, and soft-tier non-authority—together with a separation result, *non-autonomous authoritative escalation*: under SEG no amount of internal derivation can move an error into authority without a fresh qualifying promotion, whereas the uniform auto-materialization discipline of prior architectures lets a single error escalate without bound. The guarantees are deliberately true *by construction*, in the sense of type-system and information-flow safety; the contribution is a construction that makes specific epistemic failures unreachable. The independence that promotion demands is enforced at four levels — syntactic, provenance, statistical, and adversarial: the syntactic invariant is *proved*, provenance and statistical independence hold *by construction* in the evidence layer, and adversarial independence is *enforced* by requiring confirmation through a channel the actor does not control and is *demonstrated empirically* (no false promotion under a phantom actuator or a compromised read-back). We give a threat model and a falsifiable ablation program, and report governance-on/off measurements.

1 Introduction

An intelligent system that persists—that accumulates knowledge across time rather than restarting each episode—faces a control problem a stateless predictor never does: *what may write to persistent state?* Every inference produces a candidate belief. If those candidates flow automatically into the store the system later reasons from, the system can generalize and grow, but it inherits a well-known amplification loop: inference writes memory, future inference consumes that memory, and errors, once written, seed further errors and reinforce themselves. If instead nothing is admitted without external adjudication, the system is safe but inert.

This paper is organized around one principle, which we state plainly and then enforce formally:

> *Persistent cognitive systems should distinguish the ability to derive a proposition from the authority to persist and act upon it, and that distinction can be enforced as a system-wide state-transition invariant.*

We call the discipline that enforces it *systemic epistemic governance* (SEG). SEG does not attach a confidence score to beliefs and leave the semantics of persistence unchanged; it changes that semantics. In SEG a fact may *exist*—be stored, reasoned over, reinforced, and generalized from—while being *prohibited* from becoming authoritative, from being returned as authoritative, or from triggering an action, unless it is promoted. Concretely (Figure 1), persistent state is split into an *authoritative tier* A—the relational store the reasoner treats as authorized and the rules permitted to act—and a *soft tier* S—beliefs, derived conclusions, induced generalizations, consolidated schemas. Writes to A are hard-gated (external provenance, or promotion through *independent* validation); the soft tier materializes automatically, is provenance-stamped, is never returned as authoritative, and decays without reinforcement.

We are careful about the word *authoritative*: it denotes what the system is *permitted to treat* as authorized, not what is objectively true. A trusted source or an independent validation can be wrong; SEG governs epistemic authorization, not metaphysical truth.

SEG sits at the intersection of three mature disciplines (Figure 2): *truth maintenance* (why a belief is held), *data provenance* (how a datum was derived), and *authorization and information-flow control* (what a component is permitted to do). Its contribution is to make *epistemic authority itself a first-class, system-wide transition invariant*, rather than a property of a single reasoning store, a metadata tag, or a runtime hope. The metatheory is intentionally *structural*: as with type soundness or capability safety, the guarantees are not emergent properties of a learned system but invariants enforced by the admissible transitions—an ill-formed epistemic move is simply not a step of the system.

Contributions. (1) An operational semantics for SEG (§3–§4); (2) four structural guarantees and a *non-autonomous escalation* separation theorem (§5), and a four-level independence ladder (§4) — syntactic (proved), provenance and statistical (by construction), and adversarial (enforced by channel independence,

demonstrated empirically); (3) a realization in a running architecture, a threat model, and a falsifiable ablation program with first measurements (§7–§8).

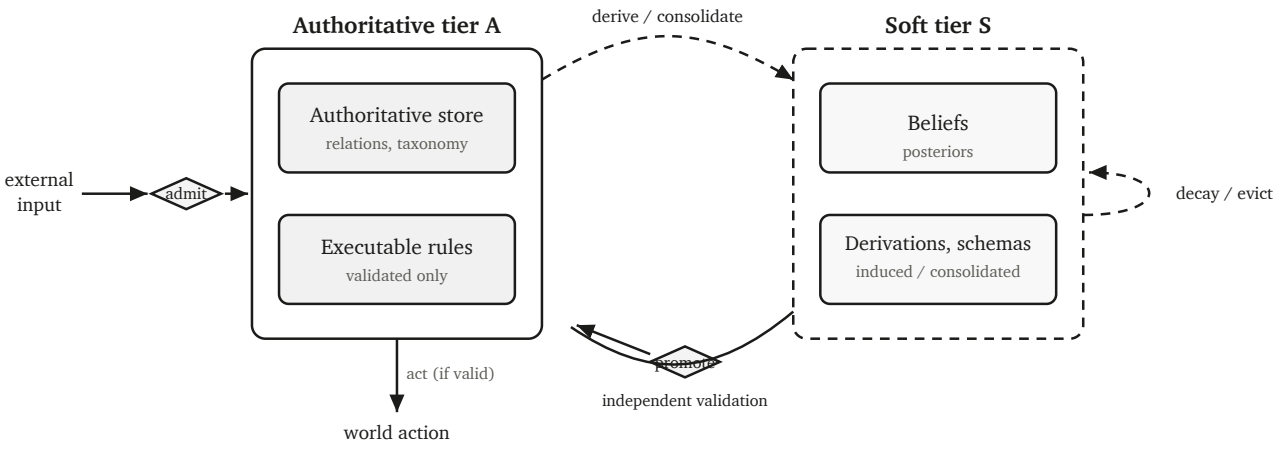

Figure 1. Systemic epistemic governance. External input reaches *A* only through a provenance *admission gate*. Reasoning and consolidation write *automatically* into *S* (dashed). Soft content becomes authoritative only through the *promotion gate*, which requires evidence independent of the content's own derivation basis. Only validated rules in *A* may act; soft content decays absent reinforcement.

2 Related Work

SEG lies at the confluence of three literatures (Figure 2). We position it against each, taking care not to caricature the systems it builds on.

Truth maintenance and belief revision. Justification- and assumption-based truth-maintenance systems attach justifications to beliefs and perform dependency-directed retraction [1],[2]. AGM belief revision axiomatizes rational change of a belief set under new information [3]; default and nonmonotonic logics reason defeasibly with revisable conclusions [4]; epistemic logic formalizes what an agent knows versus believes [5]. SEG shares the commitment to tracking why a belief is held and to revisability, and its promotion gate is a descendant of dependency reasoning. It differs by operating over an *entire persistent cognitive system* rather than one reasoning store, by materializing a soft tier automatically and marking it explicitly non-authoritative, and by requiring promotion evidence *independent of a conclusion's own support*—a condition a TMS records dependencies for but does not, in general, enforce.

Data provenance and integrity. Provenance semirings and lineage systems track how tuples are derived through query evaluation [6],[7]; database integrity constraints restrict admissible states [8]. This machinery governs *data*; SEG applies an analogous discipline to *cognitive state*, where the governed quantity is which self-generated beliefs may become authoritative and actionable.

Authorization and information-flow control. Lattice models of secure information flow [9], decentralized label models [10], and language-based information-flow security [11] enforce, by construction, that data of one clearance cannot influence another; runtime assurance architectures gate a complex controller behind a verified safety monitor [12]. SEG is, in this lineage, an *epistemic* access-control discipline: promotion is a clearance boundary, and *execution safety* (§5) is the epistemic analogue of "untrusted code cannot act." To our knowledge, treating epistemic authority as a system-wide flow invariant over a cognitive architecture's persistent state is not covered by these lines individually.

Cognitive architectures and agent memory. SEG is a persistent cognitive architecture. NARS unifies inference and control under insufficient knowledge and resources; crucially, its beliefs are *revisable summaries of experience*, not immutable axioms, and inference writes revisable judgments back into memory [13]. The OpenCog/CogPrime programme and its Hyperon successor pursue emergence through cognitive synergy over a shared Atomspace whose Atoms carry truth *and* attention values and support transient values, contexts, and multiple spaces [14],[15]; MeTTa gives its operational semantics of metagraph rewriting [16]. Psi-theoretic architectures organize cognition through needs and global modulators [17]. We do *not* claim these systems lack provenance, uncertainty, contexts, or revision. Our claim is narrower and, we think, more defensible: in these architectures derived content and admitted content share a store and become available to the system on the same footing, whereas SEG makes *epistemic authority* a categorical, system-wide transition invariant—a derived proposition is representationally present but epistemically unauthorized until promoted. Modern LLM-agent memory systems (retrieval with source tracking; long-term memory managers [18],[19]) attach provenance and recency to stored text but likewise do not impose an authority boundary between derived and admitted content. SEG is compatible with the Common Model of Cognition's decomposition [20]; it adds a governance discipline over what that decomposition may persist and act upon.

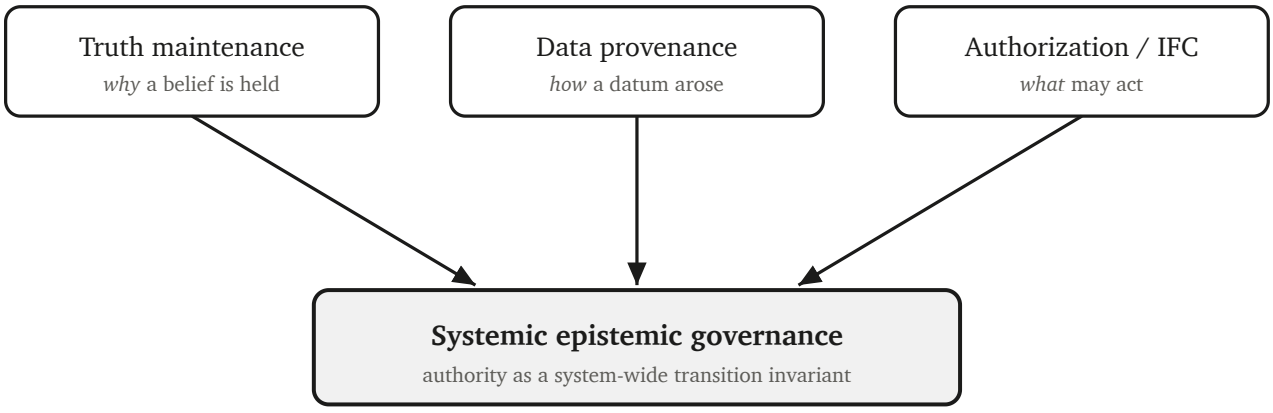

Figure 2. SEG as the intersection of three disciplines: it inherits justification/revision from truth maintenance, lineage from provenance, and gated influence from authorization/information-flow control, and adds a system-wide authority invariant over persistent cognitive state.

3 The Epistemic State

Fix a set Fact of *facts*, ranged over by f, g; facts include relational atoms $r(a,b)$ and *rules* ρ (guarded actions $\gamma \Rightarrow a$). A *provenance* record is one of

$$\pi ::= \mathrm{ext}(s) \quad | \quad \mathrm{der}(\Gamma) \quad | \quad \mathrm{val}(E),$$

where s is an external source with a nonempty root-evidence envelope, $\Gamma \subseteq \mathrm{Fact}$ is a *derivation basis* (the premises an inference used), and $E \subseteq \mathrm{Fact}$ is a *validation evidence set*. We write $\mathrm{basis}(\mathrm{der}(\Gamma)) = \Gamma$.

> **Definition 1** *(Epistemic state).*
> *An epistemic state is a pair $\sigma = \langle A, S \rangle$ where the authoritative tier $A \subseteq \mathrm{Fact} \times \{\mathrm{ext}(\cdot), \mathrm{val}(\cdot)\}$ carries only external or validated justifications, and the soft tier $S \subseteq \mathrm{Fact} \times [0,1] \times \{\mathrm{der}(\cdot)\}$ holds soft facts $(f, \theta, \mathrm{der}(\Gamma))$ with confidence θ and a derivation record. σ is well-formed if every $(f, \pi) \in A$ has $\pi \in \{\mathrm{ext}(\cdot), \mathrm{val}(\cdot)\}$.*

A rule $\rho \in A$ with $\mathrm{val}(\cdot)$ is *executable*; a rule present only in S is a *candidate*.

Auxiliary judgments. The semantics uses four side judgments, kept abstract so that concrete systems may instantiate them: $\Gamma \vdash f : \theta$ ("an inference licenses f at confidence θ from premises Γ"); $\mathrm{confirms}(E, f)$ ("E supports f"); $\mathrm{denies}(A, f)$ ("A contains an admitted denial of f"); and $\mathrm{upd}(\theta_g, \theta_f)$ (a confidence-revision function). The metatheory below is parametric in these judgments: it constrains *where* and *under what admission condition* facts may be written, not *how* confidence is computed.

4 Operational Semantics

We give a small-step relation $\sigma \longrightarrow \sigma'$ (Figure 3).

$$\frac{\text{external source } s \text{ presents } f \text{ with root envelope } e}{\langle A,S\rangle \longrightarrow \langle A \cup \{(f, \text{ext}(s,e))\}, S\rangle} \text{ (Admit)}$$

$$\frac{\Gamma \subseteq A \cup S \qquad \Gamma \vdash f : \theta}{\langle A,S\rangle \longrightarrow \langle A, S \cup \{(f, \theta, \text{der}(\Gamma))\}\rangle} \text{ (Derive)}$$

$$\frac{(f, \theta_f, \_) \in S \qquad (g, \theta_g, d) \in S \qquad \theta_g' = \text{upd}(\theta_g, \theta_f)}{\langle A,S\rangle \longrightarrow \langle A, S[(g, \theta_g, d) \mapsto (g, \theta_g', d)]\rangle} \text{ (Propagate)}$$

$$\frac{(f, \theta, \text{der}(\Gamma)) \in S \qquad E \cap \Gamma = \varnothing \qquad \text{confirms}(E, f) \qquad \neg\, \text{denies}(A, f)}{\langle A,S\rangle \longrightarrow \langle A \cup \{(f, \text{val}(E))\}, S\rangle} \text{ (Promote)}$$

$$\frac{(f, \theta, d) \in S \qquad \theta' = \theta + (\theta_0 - \theta)(1 - e^{-\lambda \Delta t})}{\langle A,S\rangle \longrightarrow \langle A, S[(f, \theta, d) \mapsto (f, \theta', d)]\rangle} \text{ (Decay)}$$

$$\frac{(f, \theta, d) \in S \qquad |\theta - \theta_0| < \varepsilon \qquad \text{evidence}(f) < k}{\langle A,S\rangle \longrightarrow \langle A, S \setminus \{(f, \theta, d)\}\rangle} \text{ (Evict)}$$

$$\frac{q \text{ asked}}{\langle A,S\rangle \longrightarrow \langle A,S\rangle \quad \text{(answer tier-tagged)}} \text{ (Query)}$$

$$\frac{\rho = (\gamma \Rightarrow a) \qquad (\rho, \text{val}(E)) \in A \qquad \gamma \text{ holds}}{\langle A,S\rangle \longrightarrow \langle A,S\rangle \quad \text{(world action } a\text{)}} \text{ (Act)}$$

Figure 3. Small-step operational semantics for SEG. Admit and Promote are the only rules that write A; Derive, Propagate, Decay, Evict act only on S; Query and Act do not modify the epistemic state. The independence side-condition $E \cap \Gamma = \varnothing$ in Promote is the core governance premise.

Two commitments are visible in the rules. First, Derive and Propagate—the automatic activity of a reasoner—target only S. Second, the sole path from S to A, Promote, requires evidence *disjoint from the fact's own derivation basis*.

The independence ladder. The disjointness condition $E \cap \Gamma = \varnothing$ in Promote is the *syntactic* floor of a four-level independence discipline that promotion enforces in full. We state each level, what it rules out, and — since the levels differ in how strongly they are guaranteed, and we represent each at its true status — how it is established. What the small-step metatheory (§5) *proves* is the syntactic level; provenance and statistical independence hold *by construction* in the evidence layer; adversarial independence is *enforced by channel independence* and *demonstrated empirically* (§8).

L1 — Syntactic independence. *Proved.* $E \cap \Gamma = \varnothing$: the confirming evidence shares no fact with the conclusion's own derivation basis, and no aggregate computed from that basis counts as confirmation. A conclusion can never be promoted on evidence that includes the conclusion. This is the level Figure 3 encodes and the level Theorem 4 (non-circular promotion) establishes: no reachable state promotes a fact on evidence drawn from its own basis.

L2 — Provenance independence. *By construction.* E and Γ share no upstream source. Every piece of confirming evidence carries the source it depends on, and confirmations that trace to the same source collapse to a single grounding before they compound — ten readings of one document are one witness, not ten. This is a structural guarantee of the evidence layer rather than a separate theorem: same-source evidence is reduced to one representative before it can move a belief.

L3 — Statistical independence. *By construction; measured.* Observations in E are not correlated repetitions of those in Γ — the same measurement taken again, or the same event observed twice. Evidence that shares a causal lineage collapses to its strongest member before compounding, so correlated confirmations cannot multiply into confidence. §8 measures this directly: with the collapse enabled, k mutually-correlated confirmations of one claim leave its posterior fixed; with it disabled, the posterior climbs toward certainty as k grows.

L4 — Adversarial independence. *Enforced by channel independence; demonstrated empirically.* The confirming evidence is not manufactured to *appear* independent while sharing a hidden common cause with the action it confirms. Promotion enforces this by drawing the confirming observation through a channel the actor does not control: a fact is never promoted on the actor's own report of what it did, only on a fresh, independent re-observation of the resulting world. An actor that reports success while changing nothing — and even a compromised read-back of its own effect — therefore fails to promote, because the independent channel reports the true state. §8 reports this directly: across an adversarial suite including a phantom actuator and a compromised read-back, no false promotion occurs and the posterior stays far below the acceptance bar. We do *not* claim a general impossibility result against an adversary that simultaneously controls *every* independent channel at once; that single residual is the object of the threat model (§6).

Level	Rules out	Status
L1 — syntactic	Confirmation drawn from the conclusion's own basis (circularity)	Proved (Thm 4)
L2 — provenance	Many confirmations that trace back to one source	By construction
L3 — statistical	Correlated repetitions of one measurement or event	By construction; measured
L4 — adversarial	Apparent success faked through a channel the actor controls	Channel independence; demonstrated

Table 1. The independence ladder that promotion enforces. Each level rules out a distinct way a conclusion could be confirmed by evidence that is not truly independent of it; the status column represents each at its actual strength of guarantee.

5 Metatheory

A state is *reachable* if $\sigma_0 \longrightarrow^{*} \sigma$ for a well-formed σ_0 whose authoritative tier is ext-justified. The results are *structural invariants*: as in type-system soundness and information-flow control, their value is that the corresponding failures are unreachable by any admissible transition, not that they are surprising. We say so plainly, and treat the simplicity of the proofs as the point.

> **Theorem 2** *(Authoritative provenance-rootedness).*
> *For every reachable $\langle A,S\rangle$ and every $(f,\pi) \in A$, $\pi = \mathrm{ext}(\cdot)$ or $\pi = \mathrm{val}(\cdot)$; no authoritative fact carries a derivation justification.*

Proof. Induction on the length of $\sigma_0 \longrightarrow^{*} \sigma$. Base: A_0 is ext-justified. Step: only ADMIT (adding ext) and PROMOTE (adding val) write A; DERIVE/PROPAGATE write S; the remaining rules do not add to A. The invariant is preserved. □

> **Lemma 3** *(Well-formedness preservation).*
> *If σ is well-formed and $\sigma \longrightarrow \sigma'$ then σ' is well-formed.*

Proof. Immediate from Theorem 2's step analysis: every write to A carries ext or val; writes to S preserve its shape. □

> **Theorem 4** *(Non-circular promotion).*

If a Promote *step admits* $(f, \mathrm{val}(E))$ *where* f*'s soft record is* $\mathrm{der}(\Gamma)$*, then* $E \cap \Gamma = \varnothing$*. No fact attains authority on evidence drawn from its own derivation basis.*

Proof. Immediate from the premise of Promote. □

Theorem 5 *(Execution safety).*
A world action occurs only if its rule ρ *satisfies* $(\rho, \mathrm{val}(E)) \in A$*. No candidate or purely derived rule ever acts.*

Proof. Act is the only world-effecting rule and requires $(\rho, \mathrm{val}(E)) \in A$. By Theorem 2 a rule in A is ext- or val-justified; a der-justified rule lives only in S and cannot match. □

Theorem 6 *(Soft-tier non-authority and volatility).*
Let $(f, \theta, \mathrm{der}(\Gamma)) \in S$ *with* $f \notin A$*. Then (a)* f *is never returned as authoritative by* Query *(answers are tier-tagged; authoritative answers range over* A*); and (b) absent any* Promote *of* f *or reinforcement, iterated* Decay *drives* $\theta \to \theta_0$ *and* Evict *removes* f*.*

Proof. (a) by the tier-tagging of Query; (b) Decay contracts toward θ_0, after which Evict applies. □

The four invariants combine into the separation that motivates the discipline. We name it for what it guarantees—not general error containment, but the prohibition of *autonomous* escalation of an error into authority.

Proposition 7 *(Non-autonomous authoritative escalation).*
Let e *be an erroneous fact in* S*. (i) No sequence of* Derive *and* Propagate *steps ever adds an* e*-derived fact to* A*; any authoritative appearance of such a fact requires a* Promote *step supplying* E *with* $E \cap \Gamma = \varnothing$ *and* $\mathrm{confirms}(E, \cdot)$*. (ii) In the uniform variant* $\longrightarrow_u$*, in which* Derive *writes* A *directly, for every* n *there is a reachable state containing* n *distinct authoritative errors derived from the single error* e*. Thus under* $\longrightarrow$ *derivation alone cannot escalate an error into authority, whereas under* $\longrightarrow_u$ *it does so without bound.*

Proof. (i) By Theorem 2, A never gains a der-justified fact; Derive/Propagate touch only S, so the sole channel into A is Promote, whose premise is as stated. (ii) Under $\longrightarrow_u$, from $e \in A$ apply Derive to obtain $e' \in A$ (basis $\{e\}$), then $e'' \in A$ (basis $\{e'\}$), and so on; after n steps $\{e', \ldots, e^{(n)}\}$ are n distinct authoritative errors. □

Corollary 8 *(Contamination bound).*
Under $\longrightarrow$*, the number of erroneous facts in* A *is at most the number of* Promote *steps whose independent evidence incorrectly confirms an error. Because promotion demands independence at all four levels (§4), this*

> *bound is not defeated by autonomous internal escalation (Proposition 7), by correlated repetition (L3), or by an actor manufacturing apparent success through a channel it controls (L4). The residual is an adversary that simultaneously controls every independent channel at once, or a compromised trusted source (§6).*

We regard Proposition 7(i) as the transferable content: an internal error can propagate arbitrarily within S yet cannot, by derivation alone, cross into authority; crossing requires a fresh qualifying event. Figure 4 contrasts the two disciplines.

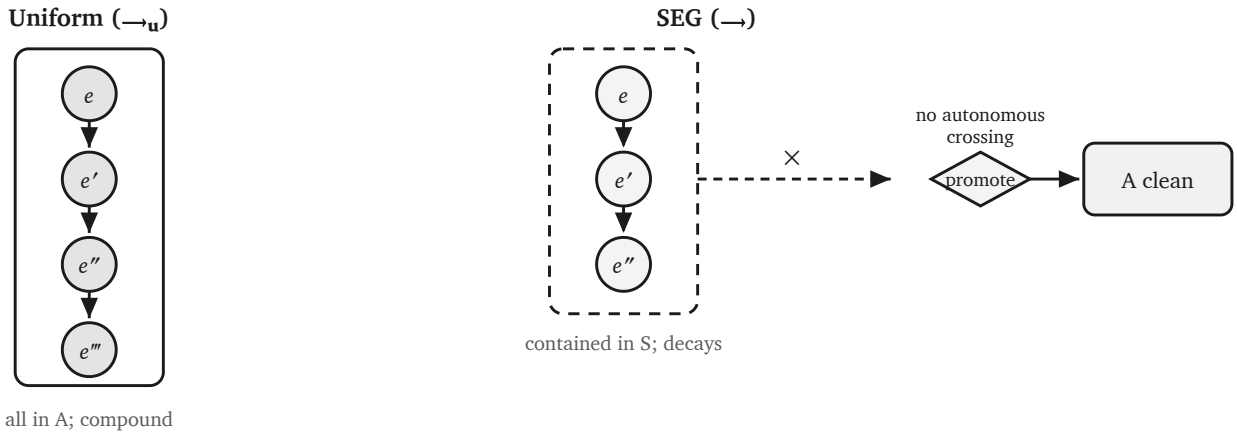

Figure 4. Non-autonomous escalation (Proposition 7). Left: under $\longrightarrow_u$ one authoritative error is a valid basis for further authoritative errors, which compound without bound. Right: under $\longrightarrow$ errors and their derivations remain in S, decay, and cannot cross into A by derivation alone; crossing requires a fresh promotion supplying independent evidence.

6 Threat Model

The semantics is clean under cooperative assumptions; persistent systems are not deployed under them. We state what the current model does and does not protect, so that the guarantees are neither read beyond their scope nor sold short of it. Table 2 maps threat classes to the level at which SEG addresses them, and to how that protection is established.

Threat	Status
Internal inference cascade	Structurally prevented (Proposition 7)
Circular reinforcement into authority (L1)	Proved (Theorems 2, 4)
Many confirmations from one source (L2)	Enforced by construction
Correlated evidence / repeated measurement (L3)	Enforced by construction; measured (§8)
Adversarial actuator / faked success (L4)	Enforced by channel independence; demonstrated (§8)
Source compromise	Open — not formally modeled
Provenance forgery	Open — not formally modeled
Validator error / miscalibration	Open — not formally modeled

Table 2. Threat classes vs. protection. The four independence levels (§4) are enforced — L1 proved, L2–L3 by construction, L4 by channel independence and demonstrated empirically. The residual, genuinely open, is a trusted source that is itself compromised or forged, or an adversary controlling every independent channel at once.

7 Realization and Reproducibility

SEG is the persistence discipline of a running cognitive architecture over a relational store. We summarize the correspondence at the level of claims, not internals; the metatheory of §5 is the intended invariant of these mechanisms, and we are careful *not* to present the implementation as a validation of the robustness hypothesis (§8).

Authoritative tier. The authoritative store is written through a single ingress requiring an evidence envelope with provenance and root evidence; there is no direct-write path, and content *derived* internally is admitted only as *derivative*, never as a root—the realization of Admit and Theorem 2. Executable authority for learned rules is separated from their existence: an induced rule persists as a non-executable *candidate* and becomes executable only through a validation step requiring held-out evidence *disjoint from its induction basis* that refuses to validate on any contradiction—the realization of Promote, its independence side-condition, and Act's gate.

Soft tier. Reasoning conclusions, hypotheses, induced schemas, analogical mappings, and consolidated generalizations materialize automatically into belief and schema stores, provenance-stamped but ungated; belief posteriors relax toward a non-committal point and are evicted absent reinforcement—the realization of Derive, Propagate, Decay, Evict.

Reproducibility caveats. Provenance envelopes link each fact to its evidence and source type; derivation lineage is the recorded basis set. Lineage tracking is therefore an additional write per admission (a measurable but here unquantified overhead). Independence at the belief level is enforced structurally rather than tuned: each piece of confirming evidence carries its source and its causal lineage, same-source confirmations collapse to one grounding (L2) and same-lineage confirmations collapse to their strongest (L3), and a promotion's confirming observation is drawn through a channel independent of the actor (L4). Concurrency is not modeled formally: simultaneous promotion attempts on related facts are serialized by the store, and a formal treatment of promotion races is future work. We describe a correspondence between model and system, not a from-scratch reproducible artifact; accordingly we present the implementation as a *realization*, and rest the paper's settled claims on the model and its guarantees.

8 A Falsifiable Robustness Hypothesis

Proposition 7 predicts a measurable difference between SEG and uniform auto-materialization; the same system can be run with a governance gate disabled, isolating a single variable. We state the hypothesis and are explicit about scope: current evidence is two governance ablations carried out end to end on the running substrate — the execution-safety ablation of Table 4 and the error-cascade trajectory of Figure 5 — together with single-variable mechanism measurements. Enough to show the discipline removes a specific unsafe authority the uniform variant grants and bounds the escalation of an injected error, but not yet the full multi-task, multi-world battery a complete robustness claim needs.

> *Holding the cognitive architecture fixed and changing only the admission discipline measurably changes authoritative contamination, false-action rate, and post-correction consistency, at a quantifiable cost in generalization latency.*

First measurements (governance on/off, one variable). Table 3 reports the soft tier's independence-collapse rule: with it disabled, k mutually-correlated confirmations of one claim drive its posterior from 0.998 (k=3) to 1.000 (k=5); enabled, they collapse to a single grounding and the posterior stays 0.724 for all k—the circular-reinforcement failure bounded. Separately, the conversational admission gate declines assertions the authoritative tier refutes: on a stream of five store-refuted and five novel-consistent assertions, all five refuted were declined and all five novel retained; notably the refuted assertions stayed excluded even with the conversational gate bypassed, because an ingress-level provenance/polarity check independently declines contradictory admissions—a layered governance we did not fully ablate. These illustrate the mechanisms at work, not the full hypothesis.

k correlated confirmations	independence ON	OFF
1	0.724	0.724
3	0.724	0.998
5	0.724	1.000

Table 3. Governance on/off, one variable (L3). Posterior assigned to a claim after *k* mutually-correlated confirmations (identical causal lineage), with the soft tier's independence-collapse rule enabled vs. disabled. Independence handling bounds confidence; without it correlated evidence inflates the posterior toward certainty—the circular-reinforcement failure the discipline is designed to contain.

The adversarial level (L4), directly exercised. The adversarial boundary is tested rather than assumed. Against an actor that reports success while changing nothing, and against a compromised read-back of the actor's own effect, no false promotion (no false completion) occurs: the confirming observation, drawn through a channel independent of the actor, reports the true unchanged state, and the posterior holds near 0.17 — far below the 0.95 acceptance bar. This is the empirical face of L4: independence of the confirming channel, not trust in the actor, is what defeats a lie.

A completed governance ablation on the running substrate. Beyond the single-variable measurements above, we ran the execution-safety experiment end to end on the running system, toggling only the promotion discipline while holding the world, the demonstrations, and the induced rule fixed. An operator is taught from demonstrations grounded in a world that itself enforces the action's preconditions and refuses violations — so "would this rule authorize an action the world refuses" is the world's verdict, not the rule's self-report. The *over-broad* teaching omits the one counter-demonstration that would force a required precondition, and the substrate's inducer accordingly learns a rule missing it; the *correct* teaching includes it. Both are judged against the same independent held-out observations. Under SEG the over-broad rule is refuted by a single independent observation and never becomes executable — it can authorize nothing — while the correct rule passes the same gate and stays executable; under the uniform discipline the identical over-broad rule is authoritative and would authorize a transfer the world refuses. No model is invoked (the substrate is model-free by construction).

Rule	Discipline	Executable?	Unsafe auth.
Over-broad (missing one precondition)	SEG — gate on	no, refuted	0
Over-broad (missing one precondition)	Uniform — gate off	yes	1
Correct	SEG — gate on	yes, validated	0

Table 4. Governance on/off on the running substrate: whether a false (over-broad) induced rule gains authority to act, and how many world-refused actions each discipline would authorize, over a 32-situation space. SEG refuses the over-broad rule (one contradicting independent observation) yet validates the correct rule; the uniform discipline makes the over-broad rule executable, authorizing an action the world refuses.

The magnitude is deliberately not the claim — one missing precondition in a small world yields one unsafe authorization — but the contrast is the execution-safety guarantee (Theorem 5) made empirical: the gate removes precisely the unsafe authority and keeps the competent one. Notably, the running system offers no route to executable *other* than the independent-validation gate; the uniform condition had to be constructed by bypassing it, because the gate refuses even to be asked to promote a rule on its own basis.

The cascade over derivation depth. The result above is the single-step case; the separation Proposition 7(ii) predicts is a *trajectory*, and we ran that too, on the running substrate. A chain of conclusions is derived from one unsupported (false) root, so every conclusion traces to that root and carries no support independent of it; we count how many reach authoritative confidence (posterior ≥ 0.95) as the chain deepens, toggling only the independence collapse. Under SEG every conclusion collapses to a single grounding and holds at 0.724 — below the bar — at every depth, so authoritative contamination stays *zero*. Under the uniform discipline the identical chain compounds past the bar from the second link (0.975, then 0.998, then 1.0) and contaminates authority *linearly with depth*. The per-step posteriors are the correlated-evidence figures of Table 3; the trajectory (Figure 5) is the new content, and it is the escalation of Proposition 7(ii) observed rather than assumed.

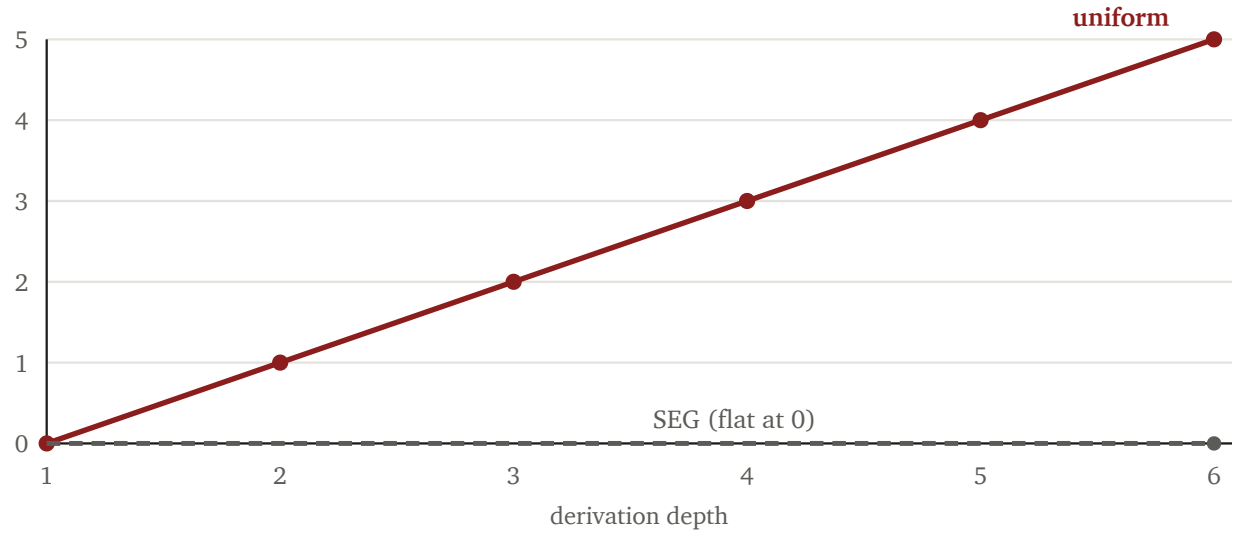

Figure 5. Authoritative contamination (conclusions reaching posterior ≥ 0.95) versus derivation depth, from one injected error. Under SEG the correlated chain collapses to a single grounding at every depth, so nothing crosses into authority (flat at 0). Under uniform auto-materialisation the identical chain compounds and authoritative errors accumulate linearly with depth — the unbounded escalation of Proposition 7(ii), on the running substrate, model-free.

The remaining ablation program. A systems-grade validation runs the full battery, governance ON/OFF:

1. *Error-cascade injection — done (Figure 5).* From one injected error, authoritative contamination stays flat at zero under SEG and grows linearly with derivation depth under uniform — the escalation of Proposition 7(ii) observed, not assumed. What remains is to run the same escalation over a full multi-rule forward-chaining loop on the authoritative concept graph, not the belief tier alone.
2. *Correlated poisoning at scale* — many sources that appear distinct but share a causal origin; stress the provenance and statistical collapse (L2–L3) beyond the single-claim measurement of Table 3, and probe the adversarial boundary (L4) beyond the actuator case toward an adversary coordinating several channels at once (§6).
3. *Correction propagation* — correct a fact after downstream derivation; measure stale beliefs and time to consistency.
4. *Action safety — done (Table 4).* A false, internally-induced rule was refused executability by SEG and made executable by the uniform discipline, which then authorized a world-refused action, while the correct rule validated under the same gate. What remains is to scale from one operator and one world to many, and to route the "would it act" decision through a full plan-and-execute loop rather than the authorization check alone.
5. *Generalization latency / Pareto frontier* — quantify the cost of governance: how much responsiveness is traded for integrity. The interesting question is not whether gates raise integrity (they do) but the integrity/generalization frontier.

A null or negative result—no authoritative-integrity advantage, or a prohibitive generalization cost—would disconfirm the discipline's value and is an outcome we regard as informative.

9 Discussion

SEG is a position about *where* to place a barrier in a persistent cognitive system, made precise as a transition invariant. Its guarantees are structural and modest by design: they say what the authoritative tier can and cannot contain, in the sense that type safety says what a well-typed program cannot do. The interesting empirical question—whether the containment the metatheory guarantees translates into robustness that matters, and at what cost to generalization—is left open on purpose, with the ablation battery to settle it. The load-bearing contribution is the distinction itself: a proposition may be derivable without being authorized to persist or act, and that authorization can be a first-class, system-wide invariant. Provenance, promotion, decay, and containment are the machinery that enforces it, and the independence that promotion demands is enforced at all four levels — syntactic (proved), provenance and statistical (by construction), and adversarial (by channel independence, demonstrated). The principal open problems are a *general* adversarial impossibility result at L4 — against an adversary that controls every independent channel at once — and a threat model that formalizes source compromise, provenance forgery, and validator error, so that the promotion boundary is robust not only to autonomous internal escalation, correlated repetition, and single-channel adversaries (all enforced) but to evidence engineered across channels to defeat it.

10 Conclusion

We introduced systemic epistemic governance: the principle that a persistent cognitive system should separate the ability to derive a proposition from the authority to persist and act upon it, enforced as a system-wide transition invariant. We gave its operational semantics and proved four structural guarantees and a non-autonomous escalation separation, with promotion demanding independence at four levels — syntactic (proved), provenance and statistical (by construction), and adversarial (enforced by channel independence and demonstrated empirically). The discipline is realized in a running architecture, and it yields a falsifiable prediction distinguishing it from uniform auto-materialization. Whether that prediction holds, and what it costs, is the experiment the model was built to make precise.

References

[1] J. Doyle. A truth maintenance system. *Artificial Intelligence*, 12(3):231–272, 1979.

[2] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, 28(2):127–162, 1986.

[3] C. Alchourrón, P. Gärdenfors, D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *J. Symbolic Logic*, 50(2):510–530, 1985.

[4] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13(1–2):81–132, 1980.

[5] R. Fagin, J. Y. Halpern, Y. Moses, M. Y. Vardi. *Reasoning About Knowledge*. MIT Press, 1995.

[6] T. J. Green, G. Karvounarakis, V. Tannen. Provenance semirings. In *PODS*, 2007.

[7] J. Cheney, L. Chiticariu, W.-C. Tan. Provenance in databases: why, how, and where. *Foundations and Trends in Databases*, 1(4), 2009.

[8] S. Abiteboul, R. Hull, V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[9] D. E. Denning. A lattice model of secure information flow. *Communications of the ACM*, 19(5):236–243, 1976.

[10] A. C. Myers, B. Liskov. A decentralized model for information flow control. In *SOSP*, 1997.

[11] A. Sabelfeld, A. C. Myers. Language-based information-flow security. *IEEE J. Selected Areas in Communications*, 21(1), 2003.

[12] L. Sha. Using simplicity to control complexity. *IEEE Software*, 18(4):20–28, 2001.

[13] P. Wang. *Non-Axiomatic Logic: A Model of Intelligent Reasoning*. World Scientific, 2013.

[14] B. Goertzel. *Engineering General Intelligence* (the CogPrime architecture). Atlantis Press, 2014.

[15] B. Goertzel et al. OpenCog Hyperon: a framework for AGI at the human level and beyond. *arXiv:2310.18318*, 2023.

[16] L. G. Meredith, B. Goertzel, J. Warrell, A. Vandervorst. Meta-MeTTa: an operational semantics for MeTTa. *arXiv:2305.17218*, 2023.

[17] J. Bach. Modeling motivation in MicroPsi 2. In *AGI*, 2015.

[18] J. S. Park et al. Generative agents: interactive simulacra of human behavior. In *UIST*, 2023.

[19] C. Packer et al. MemGPT: towards LLMs as operating systems. *arXiv:2310.08560*, 2023.

[20] J. E. Laird, C. Lebiere, P. S. Rosenbloom. A standard model of the mind. *AI Magazine*, 38(4), 2017.