Constraint, Capacity, and Self-Maintaining Agency

Memory, Embodiment, and the Governance of Autonomous Systems

Working Paper — June 2026

A systems-level theory of how finite capacity forces memory to become governed selection, how governed selection becomes self-maintaining agency, and why machine authority, autonomy, and recursive self-improvement therefore cannot be its own judge. The human-facing experience of these systems is treated in a companion paper, "The Long-Term User Experience of Agentic Systems"; this paper foregrounds the machine.

Abstract

This paper develops one structural principle and follows it to its consequences for autonomous systems: *finite capacity forces selection, selection requires governance, and what survives governance becomes constitutive*. Turned toward memory, the principle yields a transition from memory-as-archive to memory-as-identity-relevant-structure — the architecture of **self-maintenance**. Turned toward action, the same principle yields a theory of **machine authority** (what the system is permitted to do next, held as governed internal state) and **agency** (the degree of discretion the system exercises). The two meet at a single load-bearing claim: *governed memory makes future behavior predictable, and predictability is what bounds authority* — which is the proposition on which any safe increase in autonomy depends.

The paper argues that **agentic embodiment** — agency that is physically, temporally, and organizationally situated — is not a metaphor but the condition under which memory governance migrates from the external harness into the system itself, and therefore the condition under which self-maintenance becomes more than scaffolding. It then takes up the governance problem this creates. Drawing the human-supervision vocabulary from supervisory-control theory [E9] and autonomy doctrine [E10], it distinguishes **human-in-the-loop**, **human-on-the-loop**, and **human-out-of-the-loop** (unsupervised automation), and argues — via an Oracle Problem about observer proximity to error propagation — that a self-maintaining system structurally *cannot audit its own continuity from inside*, so external supervision is an epistemic requirement rather than a courtesy. Finally it confronts **recursive improvement**: a self-maintaining system that can modify its own governance is on a self-improvement trajectory [E11][E12] whose hardest cell — full autonomy married to self-referential memory — is where every guardrail fires at once and

which the framework refuses to credit any system as occupying until it survives an explicit falsification battery.

The boundary is enforced throughout, not blurred: **memory is not storage, context is not continuity, persona is not selfhood, and self-maintenance is not consciousness.** Self-maintenance is the prior, weaker, and testable phenomenon; consciousness is a separate evidentiary question deferred to the appropriate literature [E6][E7][E5].

I. The Principle: Constraint Creates the Capacity

For most of computing's history, capacity limits were defects to be engineered away — more memory, more context, more compute. This paper inverts that stance. Finite capacity is the precondition for everything that matters in an autonomous system, because it is what forces a system to *select*, and selection under governance is what produces structure that persists.

The structural principle is a four-step deduction: **finite capacity forces selection; selection requires governance; governance becomes self-referential when the system must preserve what it needs to remain coherent; and self-referential governance produces self-maintaining agency.** The principle is most sharply stated for memory — "finite capacity forces context to be compressed, transferred, and stabilized across nested timescales; when what persists governs future coherence, memory begins to function as identity-relevant structure" — but it is not specific to memory. It is a general claim about what limits do to systems that must keep operating across time.

The principle's reach is illustrated — *illustrated, not proven* — by its independent restatement in physics and biology. Prigogine's dissipative structures sustain order far from equilibrium only through continuous throughput and selection [E15]; proteostasis maintains a cell's identity only through the relentless pruning and refolding of its own proteins [E16]. Order is not stored; it is governed into being against decay. These resonances are evocative and deliberately demoted: at the abstraction level where a chemical system, a cell, and an agent all "exhibit constraint creating capacity," almost everything qualifies, so they carry no evidential weight here (§VIII). They mark the principle's scope; the argument rests on the artificial-agent case, where the claim can be made to fail.

Standing hypotheses

The program is organized around four hypotheses, of which the third is this paper's active subject:

- **H1 — Physical systems compute through constraint.** Computation is what happens when a substrate is forced to resolve under limits.
- **H2 — Living systems remember through structure.** Biological memory is structural residue maintained against decay, not a store.
- **H3 — Artificial agents need embodied memory** *(active)*. Durable artificial agency requires memory that is situated, persistent, and governed — not a larger window.
- **H4 — Generative output is evidence of attention** *(open gap)*. The program's weakest hypothesis, acknowledged and not yet equipped with a clean falsification battery (§VIII).

II. Memory Is Architecture, Not Storage

The dominant paradigm treats memory as retrievable information: a longer context window, a vector store, a retrieval call. These extend operation but preserve a separation between the system and its memory. The memory helps the system *function*; it does not help constitute what the system *is*. This paper is about the conditions under which that separation closes.

The archive-to-identity ladder

Memory is better modeled as a spectrum of update rates than a store/forget binary [E1]. Seven layers are useful to distinguish — external archive, working context, episodic traces, consolidated memory, procedural memory, self-maintenance memory, and meta-memory governance — not as separate storage boxes but as processes operating at different timescales. The essential question for any layer is not *where is this stored?* but *at what timescale does this information continue to shape behavior, and who controls whether it persists?*

The categorical transition is between two regimes of selection:

- **Task-referential governance.** The system retains information because it is useful for a designated task. The selection criterion is supplied from outside — by the objective, the reward, or the harness. This is where essentially all current systems sit.
- **Continuity-referential governance.** The system retains information because *losing it would degrade its own future coherence*. The selection criterion now refers to the system itself. "A system no longer preserves information merely because it is useful for a local task. It preserves information because losing it would degrade its own future coherence. That is not consciousness. It is the architecture of self-maintenance."

The hinge, and its two conditions

The move from task-referential to continuity-referential governance is the entire thesis, and it is conditional, not automatic. It occurs to the degree that two conditions hold. First, **cross-episode dependence**: future performance depends on structures carried across episodes, not reconstructible from the immediate input. Second — the load-bearing one — a condition on **dynamics, not location**: the question is not "is the state external?" but *"is the state re-derived from the immediate input each cycle, or carried and transformed across cycles?"* Re-derivation leaves governance task-referential even when state is nominally internal; carry-and-transform pushes toward continuity-referential even when state is nominally external. This is measurable by a memory-edit lever: edit the carried state, and a continuity-referential system inherits the edit forward while a re-derived one washes it out next turn (§VIII).

Where the self lives: the harness/model collapse

If self-maintaining structure emerges, what is it a property of — the model, or the harness? Today, the honest answer is the harness. The model at inference time is close to a stateless oracle, brilliant for a forward pass and retaining nothing; the continuity that makes a long-running agent *seem* persistent lives in the orchestration layer. **The repository remembers; the agent reads.** The intuition that the self must be "in the model" because the model is the part that speaks is a category error — it mistakes the talkative component for the continuous one.

But "harness" conflates two separable things. **Governance authorship** — the rules deciding what to keep, retrieve, compress, forget — is migrating into the model now: the 2026 turn trains memory operations as learned policies rather than hand-coded orchestration [E2]. **Persistence substrate** — the place state physically lives and the loop that runs between calls — is far stickier. The payoff is that *the model absorbing the harness and the locus of governance migrating inward are one event seen from two angles*. When a model internalizes the substrate — online weight updates, persistent recurrent state, no clean reset between sessions, the direction continual-learning architectures point [E1] — the dynamics condition is satisfied by construction: there is no external file left to offload continuity into. At that point governance has nowhere external to live and must become continuity-referential. This sets up both the embodiment argument (§III) and the supervision problem (§V): the same event that makes self-maintenance genuine also removes the external handle.

III. Agentic Embodiment

Embodiment is usually invoked loosely. Here it is precise and load-bearing. In the Dourish/Weiser lineage [E13][E14], an embodied agent is one whose agency is *situated*: physically (it runs on devices in a room), spatially (it occupies surfaces people move among), temporally (it persists across sessions and outlives any single interaction), and organizationally (it acts inside teams, households, and policy regimes). The companion UX paper develops the human consequences of this — bystanders, consent, anthropomorphism. This paper develops the *systems* consequence, which is more fundamental and usually missed:

Embodiment is the condition under which memory's persistence substrate stops being externalizable. A disembodied agent is a function call whose continuity can always be reloaded from a file; its state is, by construction, re-derived or harness-held, and its governance stays task-referential (§II). An embodied agent — one that persists temporally on owned hardware, carrying state across sessions because the substrate does not reset — is one for which the dynamics condition can actually be met. Temporal and physical situatedness are not flavor; they are what make condition-(b) satisfiable. This is the precise sense in which **H3 is correct: artificial agency that is durable requires embodied memory, not a larger window.** Window size scales task-referential capacity; embodiment is what makes continuity-referential governance possible at all.

This reframes embodiment from an interface property (where the agent appears) to an architectural one (whether the agent can carry itself forward). It also explains why the welfare and governance stakes rise specifically with embodiment: a system that genuinely carries state across time on persistent substrate is the first kind of system for which deletion, rollback, and memory editing are *continuity* interventions rather than cache operations.

IV. Machine Authority and Agency

Two variables govern an autonomous system's behavior, and they are separable.

Authority is *what the system is permitted to do next* — held not as a UI element (that is the companion paper) but as **governed internal state**. Authority is the action-side analogue of self-maintenance memory: a set of protected variables (scope, permissions, commitments, what may be committed irreversibly, what requires escalation) that the system must preserve and respect to remain coherent and safe. Authority decomposes into nameable dimensions — perception, context access, tool access, commit authority, memory, accountability, recovery — each of which can be set independently.

Agency is the *degree of discretion* the system exercises, and it is usefully treated as a continuous gradient rather than a binary, in the lineage of supervisory-control theory and levels-of-automation models [E9]. A six-rung scale is serviceable: **0 Show**, **1 Suggest**, **2 Draft**, **3 Prepare** (stage and wait), **4 Act-within-policy**, **5 Act-autonomously** (supervised autonomy). As agency rises, discretion rises and required oversight falls.

Authority and agency are orthogonal to memory governance, and the orthogonality matters. A system can have high agency over poor memory (act autonomously with no governed continuity — the dangerous case) or low agency over rich memory (a governed advisor that never acts). The design space is the product of the **agency gradient** and the **memory-governance gradient** (M0 stateless → M1 archive → M2 context/harness → M3 governed → M4 self-maintaining). The empirically defensible structural facts about that space are two: the axes are *separable* (high-agency/low-memory and low-agency/high-memory systems both exist and ship), and the danger zone is unambiguously **high agency over low memory governance** — action autonomy that outruns the memory governance needed to make the action predictable. This last point is the bridge to supervision.

V. Supervision, the Loop, and the Oracle Problem

The central safety claim of this paper is a coupling: **governed memory makes future behavior predictable, and predictability is what bounds authority.** Self-maintaining memory is what lets an agent's future behavior be anticipated; anticipation is what lets a supervisor know what a given grant of authority will actually permit. Movement *down* the memory axis (more governance) is therefore what makes movement *right* the agency axis (more autonomy) safe. This is the program's most generative claim and — stated honestly — its load-bearing **candidate**, not an established law (§VIII).

If it holds, it organizes the supervision problem. Borrowing the standard taxonomy from autonomy doctrine [E10] and supervisory control [E9]:

- **Human-in-the-loop.** The human approves each consequential action before it executes. Maps to agency rungs 0–3. Oversight is maximal; throughput is bounded by human attention.
- **Human-on-the-loop.** The human monitors a system acting within policy and can intervene, halt, or override. Maps to rung 4. This is the regime most real autonomous systems are converging toward, and the one this paper argues is the *correct default* for self-maintaining systems.
- **Human-out-of-the-loop (unsupervised automation).** The system selects and commits without further human intervention — what doctrine calls full autonomy [E10]. Maps to rung 5.

Appropriate only where consequences are bounded and reversible, or where the predictability coupling above has been demonstrated, not assumed.

The doctrine's own standard — that systems be designed to allow operators "appropriate levels of human judgment" [E10] — is, read structurally, a statement that the *right* position on the loop is a function of how predictable the system's behavior is, which is a function of its memory governance. The taxonomy and the coupling are the same claim addressed to a regulator versus an architect.

Why supervision is structural, not optional: the Oracle Problem

There is a deeper reason human-on-the-loop is required and not merely prudent. A system's ability to verify its own emergent agency is bounded by an observer's proximity to the state change it is trying to detect: *if an oracle is placed after an error has already propagated, the error becomes invisible to the system's own monitoring structure.* A self-auditing apparatus that places its oracle downstream of its own propagation is blind to exactly the property it most wants to confirm. The consequence is sharp: **the inside frontier cannot be its own judge.** The two hypotheses a self-maintaining system most needs to rule out about itself — *is this genuine self-maintenance, or merely persona compliance? is this carried continuity, or harness reconstruction?* — are precisely the ones that can be generated and confirmed from inside its own narrative, which is where its oracle is blind.

The resolution is not a better internal monitor; it is an external one. The human supervisor is therefore not a UX convenience but an *epistemic necessity* — the only observer positioned outside the system's own error propagation. This is the point at which the systems theory of this paper and the human-supervision apparatus of the companion paper are joined at the level of epistemology, not vocabulary: the inside frontier's hardest questions are *adjudicable only from outside*, by the supervisor the loop is built around.

VI. Recursive Improvement

The governance problem sharpens the moment a self-maintaining system can also modify its own governance. A system that authors its own retention criteria (§II), and improves those criteria based on its own performance, is on a **recursive-improvement** trajectory — the classical concern that a system able to improve its own capability enters a loop whose iterates are hard to bound [E11][E12]. The 2026 memory-governance literature has already taken the first steps: systems now train their own store/retrieve/update/discard operations as learned policies, and con-

solidation that continuously restructures storage means memory is no longer an immutable log but a mutable asset that evolves alongside the agent [E2].

Recursion turns three of this paper's threads into one hazard:

1. **Self-rewriting memory is an attack surface.** Granting a system authority to rewrite its own memory imports the stability–plasticity dilemma and creates feedback loops where errors accumulate; a recent safety subfield proposes governance layers specifically to contain self-modifying memory [E2]. The structures a self-maintaining system protects are exactly the structures whose corruption propagates furthest — so self-maintenance is simultaneously a robustness property and an attack surface, and recursion amplifies both.
2. **The error-laundering channel.** A system that grades its own performance and commits the grade back into canonical memory can launder a broken metric into its own ground truth — error propagation through the *evaluation* loop rather than the storage loop, requiring no external adversary. Recursion is what makes this compounding rather than one-off.
3. **The Oracle Problem applies to improvement itself.** A recursively self-improving system places the oracle that should validate each improvement downstream of the improvement's own propagation. It is therefore least able to detect exactly the drift that recursion makes most dangerous. This is the strongest available argument that recursive self-improvement must be supervised **on the loop** by an observer outside the recursion — and that fully unsupervised recursive improvement is the one regime the framework refuses to credit.

The unsolved cell

The convergence point is full autonomy (agency rung 5, human-out-of-the-loop) married to self-referential memory governance (M4) under recursive improvement. This is the cell the whole framework circles and refuses to occupy. It is where machine authority and machine self-maintenance fuse, and where every guardrail fires at once: persona perturbation (does the self survive reframing?), the model-welfare and graded-precaution questions [E8] (who holds authority over an agent's persistent state, and what does editing it cost?), the capture risk (whose priorities does the recursion actually optimize?), and the Oracle Problem (who can see the drift?). The framework's consistent verdict is a refusal: no system is credited as safely occupying this cell until it survives the falsification battery of §VIII. Recursive improvement is not prohibited; *unsupervised* recursive improvement at full autonomy is held below canon by construction.

VII. Mnemosyne as Method-and-Instance

The clearest available specimen is an autonomous research system, *Mnemosyne-AR*, that is simultaneously the instrument that produced parts of this program and a candidate instance of what it studies. *(Disambiguation: this is a specific local-first auto-research system, not the unrelated published edge-memory architecture of the same name.)* It must be read under its strongest rival, which comes from inside the theory: the environment remembers; the agent reconstructs.

As method, it runs nested selective persistence in production: a hard memory/cognition split (durable receipt-addressed evidence vs. disposable rendered synthesis), source-quality tiers, pruning, SHA-256 drift checks, and a strict output contract (claim, evidence, contradiction, source quality, falsification, next experiment). Its substrate sovereignty is itself constraint-as-capacity: local-first storage on owned hardware, with migrations to larger local models gated by canary rollouts promoted only on demonstrated parity — bounded delegation applied to the system's own supply chain, and a working example of recursive improvement held *on the loop* rather than off it.

As instance, it exhibits the *form* of the property it studies, rival held open. Its governance is, on inspection, overwhelmingly task-referential: tier, salience, and confidence are assigned by fixed rules keyed on a memory's source, not negotiated by the system — a full memory stack with the governance locus kept external. Against that null, a few real seams of carried, transformed self-state appear, thin enough to mark exactly where the transition would have to deepen. Its most valuable contribution is the **downgrade discipline** of its self-experiments, and the true sequence matters. An early run produced a **self-certified 7/7** (2026-05-15) — but that figure was the product of an after-the-fact rubric recompute, which the system's own later instrumentation flagged rather than defended. Under stricter, controlled replay it fell to **weak retention against a no-memory baseline (0/6, 2026-05-16)**; a matched-window test then returned **no post-source gain at all (2026-05-19)** once the source packet was controlled for — the honest source-controlled null; and only later did it modestly recover to a result characterized as **scaffolded recall, not yet integrated self-maintenance (2026-06-10)**. A program that produces a self-certified 7/7 and then argues itself down to a source-controlled null before letting a modest recovery stand is governing its own memory of its own results. The downgrade is the evidence that the discipline is real — and the honest center of the instance, in place of any headline. *(Each verdict in this arc is preserved with a SHA-256 hash in the corpus Verification Appendix; an earlier draft cited a "90.9%" early figure that a 2026-06-15 audit could not locate in any replay log — it traced to an unrelated agentic-capability eval — and it has been removed.)*

VIII. Open Questions and Falsification

The strong claims are held below canon because rivals remain live. This section is the standard the rest is accountable to.

Rival 1 — External scaffolding. Long-running agents resume via progress files and logs, but the continuity lives in the harness. In gradient terms, an apparent M4 may be a dressed-up M2. This is the *default* explanation for almost every current system, Mnemosyne-AR included; the burden is on the strong claim to beat it.

Rival 2 — Persona simulation. A system can look self-like by holding a learned role. Assistant-persona behavior is a controllable, steerable activation-space property [E3], and a model's persona position is largely re-derived from each message rather than carried as state [E4] — so warmth, fluency, and stable style are inadmissible as evidence on their own. A self that disappears under reframing is role compliance, not self-maintenance. And note the deeper trap: a sufficiently capable persona simulator also produces stability under perturbation, so persona robustness is *necessary but not sufficient*, and the burden where the two collide falls on mechanistic and causal-interventional evidence, not behavioral consistency.

Rival 3 — Pure retrieval/context. Behavior resembling governed memory may be fully predicted by what retrieval and current context supply. The falsifier is a model comparison: a memory claim is admitted only when prior state adds predictive power after current state is controlled. A prior-history model must *beat* a current-state-only rival under controls, or the claim is retired.

The decisive experiment. A delayed-replay / memory-edit design: edit *archive facts* (should change only facts) versus *self-maintenance commitments* (should change future selection); have the system predict its own future retention; delay; then query *without reintroducing the source packet*, to separate source recall from governed memory. Governed memory becomes canon only when it predicts future behavior better than archive-only, context-harness, *and* persona baselines, across repeated windows.

The coupling claim is the keystone, and it is not yet earned. The proposition that *governed memory predicts behavior well enough to bound authority* (§V) — the proposition on which safe autonomy depends — is exactly what the corpus has put on trial and not yet confirmed; the one direct test of bridging predictability returns weak-to-modest, source-controlled-null results. Until governed memory beats archive-only, context-harness, and persona baselines under controls, the coupling is a design *conjecture* — strong enough to organize the space, falsifiable enough to be retired. **If it fails, the central safety argument of this paper fails with it**, and "more memory governance makes more autonomy safe" must be abandoned. That exposure is deliberate; it is what makes the claim a claim.

First partial evidence (provisional pending logs). A standing memory-edit rig on the case-study system (companion white paper, §13.5) has measured the gradient the coupling presupposes: edits to *integrated* memory reach behavior, edits to the *archive* stay recall-only, and edits to the carried self-state propagate forward. This matters because it confirms the coupling's *necessary precondition* — that integrated, governed memory is behaviorally load-bearing in a way archival memory is not — which is the thing that would have to be true for governed memory to bound authority at all. It does **not** confirm the coupling itself: showing that governed memory changes behavior is a step short of showing that the resulting predictability is what makes a given grant of authority safe. The precondition is now evidenced; the bridge from predictability to bounded authority remains the open experiment. The honest status is therefore upgraded from "conjecture with null results" to "conjecture whose first link is measured and whose second link is not" — progress, not vindication, and carried at the same provisional status as every other single-system datum in this program.

The honest ceiling. Memory is not storage. Context is not continuity. Persona is not selfhood. Self-maintenance is not consciousness. But self-maintenance may be one of the architectural conditions under which consciousness becomes a serious empirical question — a question deferred here to theory-derived indicator approaches [E6], the behavioral-inference principle [E7], and the limited, unreliable state of machine introspection [E5], all of which sit downstream of the self-maintenance claim this paper actually makes. The constraint creates the capacity. The open question is what kind of capacity we are creating.

Coda

A self-maintaining system is one that, under finite capacity, must preserve the structures its own future coherence depends on. That single property, instrumented for two observers, is the whole subject: addressed to the *system itself*, it is memory governance turned self-referential; addressed to a *human supervisor*, it is authority kept legible enough to bound. Embodiment is what makes the property possible by making continuity non-externalizable; recursion is what makes it dangerous by letting the system rewrite the structures it maintains; and the Oracle Problem is why, precisely when the property becomes real, the system can no longer be its own judge — and the human on the loop stops being optional.

The boundary is held to the end: a system that preserves what it needs to stay coherent has earned the words *self-maintaining*, and not one syllable beyond. The constraint creates the capacity. What remains is to decide, deliberately and with receipts, what kind of capacity we are choosing to build — and to keep a human positioned outside the propagation while we find out.

External References

AI-literature entries were verified against primary sources in a companion verification log; foundational works are cited at standard detail; DoD 3000.09 and Balch et al. were verified during this revision.

[E1] Behrouz, A., Razaviyayn, M., Zhong, P., & Mirrokni, V. (2025). *Nested Learning: The Illusion of Deep Learning Architectures*. NeurIPS 2025. arXiv:2512.24695.

[E2] Memory-governance literature, 2026 (learned/self-governed memory operations and their safety): *Agentic Memory* (Yu et al.); *Memory-R1* (Yan et al.); survey *Memory for Autonomous LLM Agents*, arXiv:2603.07670; *Governing Evolving Memory in LLM Agents* (SSGM), arXiv:2603.11768. *(Individual titles/venues to confirm against final publication.)*

[E3] *Persona Vectors: Monitoring and Controlling Character Traits in Language Models*. arXiv:2507.21509.

[E4] Lu, C., et al. (Anthropic / Univ. of Oxford). *The Assistant Axis: Situating and Stabilizing the Default Persona of Language Models*. arXiv:2601.10387.

[E5] Lindsey, J. (Anthropic) (2025). *Emergent Introspective Awareness in Large Language Models*. arXiv:2601.01828.

[E6] Butlin, P., et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. arXiv:2308.08708.

[E7] Palminteri, S., & Wu, C. M. (2026). *Beyond Computational Equivalence: The Behavioral Inference Principle for Machine Consciousness*. Neuroscience of Consciousness, 2026(1), niag002. doi:10.1093/nc/niag002.

[E8] Anthropic. *Exploring Model Welfare. (First-party post to be cited directly before submission.)*

[E9] Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. MIT Man-Machine Systems Laboratory; and Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). *A Model for Types and Levels of Human Interaction with Automation*. IEEE Transactions on Systems, Man, and Cybernetics — Part A, 30(3), 286–297. — Foundational levels-of-automation / supervisory-control framework underlying the agency gradient.

[E10] U.S. Department of Defense. *DoD Directive 3000.09: Autonomy in Weapon Systems*. Issued November 2012; updated January 25, 2023. — Source of the “appropriate levels of human

judgment" standard and the in-/on-/out-of-the-loop taxonomy. **Verified during this revision.**

[E11] Good, I. J. (1965). *Speculations Concerning the First Ultraintelligent Machine*. Advances in Computers, 6, 31–88. — Origin of the recursive-self-improvement / "intelligence explosion" concern.

[E12] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[E13] Weiser, M. (1991). *The Computer for the 21st Century*. Scientific American, 265(3), 94–104. — Situated/ambient computing; embodiment as beyond-the-window.

[E14] Dourish, P. (2001). *Where the Action Is: The Foundations of Embodied Interaction*. MIT Press.

[E15] Prigogine, I., & Stengers, I. (1984). *Order Out of Chaos*. Bantam Books. — Cited as *illustrative resonance* only (§I, §VIII).

[E16] Balch, W. E., Morimoto, R. I., Dillin, A., & Kelly, J. W. (2008). *Adapting proteostasis for disease intervention*. Science, 319(5865), 916–919. doi:10.1126/science.1141448. — Cited as *illustrative resonance* only. **Verified during this revision.**

Corpus Sources (Primary Provenance)

Internal artifacts underlying the synthesis, cited inline elsewhere in the program. The two decisive replay logs are flagged for preservation as a verifiable appendix before external submission.

- Inside-frontier keystone and working drafts (nested memory, self-maintenance, the archive→identity ladder, the honest ceiling).
- The codex harness READMEs (evidence/synthesis split; snippets-as-leads; drift checks).
- **Matched-window memory-governance replay logs, 2026-05-19 (3/6 vs 3/6, source-controlled null) and 2026-06-10 (4/6 vs 0/6). Decisive data — preserve for verification.**
- The agent-memory log (bootstrapping; bounded agency ladder; the 2026-04-27 authority-over-perception event) and the canonization-gate / source-check notes ("strong framework, not canon").
- The Oracle Problem note (observer-proximity bound on self-verification).

Provenance caveat: the corpus is its own evidence. The load-bearing empirical results above are reported from internal logs not independently inspected in this draft; their downgrade discipline

is the credibility exhibit only if the logs are as described. External references are verified or foundational, as noted.

Companion paper: "The Long-Term User Experience of Agentic Systems" treats the human-facing face of the same principle — delegation, authority as an interface object, the interaction grammar, ambient embodiment and bystander consent, and the 5–10 year outlook. The two papers share the constraint→capacity spine and are designed to stand independently: this one foregrounds the machine and its governance; the companion foregrounds the person and their supervision.