The Agency-Semantics Spine of Distinction Theory

Attention, Value, Goal, Meaning, and Action under Finite Capacity

Yining Wu
Independent Researcher
yining.wu@alumni.upenn.edu

Version v1.0 – May 2026

Abstract

FDS-M0 develops the agency-semantics spine of Distinction Theory. Building on the formal core of Active Finite Distinction Systems, it treats attention, value, goal, meaning, and agency as finite-system roles in boundary maintenance under limited capacity. A distinction becomes attended when admitted into a finite update channel; valuable when it has causal boundary-gradient relevance; goal-relevant when the value ranking is stabilized into a policy orientation; meaningful when compressed into a task-sufficient actionable quotient; and agentic when updates or actions causally affect future boundary-maintenance loss. The paper does not claim to solve consciousness, reduce semantics to physics, or assign strong agency to any system merely because it processes symbols. It provides a conservative bridge for analyzing biological, cognitive, artificial, and social systems as active finite systems that admit, rank, compress, verify, externalize, and act on distinctions under capacity and resource constraints. A deterministic synthetic normal-form model illustrates nonlinear attention gating, semantic capacity deficit, policy-preserving quotients, goal stability, boundary-sensitive agency classification, misalignment audits, and epistemic pollution as verification-bandwidth saturation.

Keywords: agency; semantics; attention; value; goal; meaning; finite capacity; boundary maintenance; active finite distinction systems; artificial agency; misalignment; verification load; epistemic pollution.

Epistemic Notice and Scope

This manuscript is a spine paper, not a completed theory of mind or a blueprint for artificial agents. It does not claim that subjective experience reduces to information capacity, that all semantics is purely functional, that moral value reduces to survival, that current language models possess strong FDS agency, or that linguistic and social meaning can be replaced by a single information-theoretic quantity. Its claim is narrower: once a system is modeled as an active finite distinction system with a boundary, memory, update rule, action space, finite capacity, and boundary-maintenance loss, several agency-semantics notions acquire operational roles.

The paper follows the layered discipline of the FDS formal core. The FDS Core separates formal definitions, physical bridge assumptions, normal-form dynamics, and quarantined applications; it also states that domain applications must specify their boundary, memory, update rule, task variable, capacity, resource budget, perturbation family, invariant quotient, and falsification condition [1]. M0 uses that discipline for agency and semantics. A failure of a cognitive, artificial, social, or linguistic mapping may demote the mapping without refuting the formal FDS core.

1 Introduction

A finite system does not encounter the world as an unlimited field of meaning. It encounters more possible distinctions than it can admit, maintain, rank, compress, verify, or act upon. A flash in peripheral vision, a new word in a conversation, a rising temperature in a cell, an alarm in a control room, a legal right in a society, and a prompt in an artificial system are not semantically equivalent merely because they are all signals. They become operationally significant only if some system can admit them into an update channel, maintain them as records, rank their boundary relevance, connect them to possible actions, and preserve enough structure for future use.

The FDS core defines an active finite distinction system as a tuple

$$S = (X, E, B, M, Y, A, U, \pi, \ell, \Phi, \mathcal{P}, \tau), \tag{1}$$

where X is internal state, E environment, B boundary, M memory/model space, Y observation channel, A action space, U update map, π finite projection, ℓ boundary-maintenance loss, Φ resource budget, $\mathcal{P}$ perturbation/pruning family, and τ update timescale [1]. M0 asks what follows for agency and semantics when distinctions must pass through this finite architecture.

The dependency skeleton is

$$\text{distinction} \to \text{record} \to \text{attention} \to \text{value} \to \text{goal} \to \text{meaning} \to \text{agency} \to \text{culture}. \tag{2}$$

This is not a temporal sequence and not an evolutionary law. It is an operational dependency map. A distinction separates alternatives. If maintained, it becomes a record. If admitted into a finite update channel, it becomes attended. If ranked by causal boundary relevance, it becomes valuable relative to a task. If stabilized over update windows, it becomes goal-relevant. If compressed while preserving action, prediction, verification, or coordination affordances, it becomes meaningful. If connected to updates that causally affect future boundary loss, it participates in agency. When externalized and shared, such distinctions become culture or institutional infrastructure.

1.1 Why a spine paper is needed

The preceding FDS sequence developed finite record formation, register time, capacity overflow, boundary-maintaining self-organization, and finite-memory boundary-maintenance channels [2–6]. The physical-accounting ladder is background to this paper rather than a direct premise. These papers describe what finite systems can represent, maintain, update, erase, externalize, and lose. M0 shifts the question: how do maintained distinctions become significant for action? Without such a spine, later papers on attention, value, meaning, AI alignment, institutions, science, law, education, and civilization memory risk becoming disconnected applications. With it, these domains can be treated as different instantiations of finite distinction admission, ranking, compression, verification, and boundary update.

1.2 What this paper does not claim

M0 does not solve consciousness. It does not decide the metaphysics of intentionality. It does not claim that value is identical to survival or that ethics can be derived from boundary maintenance. It does not claim that meaning is only private, nor that language and institutions are irrelevant. It does not assign strong FDS agency to a bare model, static database, or passive sensor. It does not give a training recipe, benchmark, routing algorithm, memory architecture, or foundation-agent design. It provides a public conceptual spine, not an engineering blueprint.

Table 1: Claim-status summary for FDS-M0. The table is an audit device: several entries are bridge definitions or downstream interfaces, not established empirical claims.

Claim ID	Tier	Claim	Failure or demotion condition
M0-001	Formal bridge	Attention is capacity-limited distinction admission into an update channel.	Attention-like selection occurs without any capacity-limited admission, update gating, or priority constraint under the specified mapping.
M0-002	Bridge definition	Value is causal boundary-gradient relevance under finite capacity.	Valuation systematically fails to correlate with future boundary loss, task success, or resource relevance under a valid mapping.
M0-003	Operational bridge	Goals are stabilized value rankings coupled to policies across update windows.	Goal-like behavior persists without memory, ranking, policy stabilization, or action direction over time.
M0-004	Bridge definition	Meaning is actionable compressed distinction preserved by a task-sufficient quotient.	Compressed representations guide no action, prediction, verification, coordination, or boundary maintenance.
M0-005	Definitional / empirical	Strong FDS agency requires updates or actions that causally affect future boundary loss.	A system with no causal update effect on future boundary loss qualifies as a strong agent under the same criteria.
M0-006	Operational bridge	Self-verifying agency requires internal verification of whether actions reduce loss.	A system is classified as self-verifying despite relying entirely on an external host for verification.
M0-007	Empirical bridge	Misalignment is divergence between host and delegate action effects on boundary loss.	Divergent objectives do not produce divergent finite-difference action effects under audit.
M0-008	Downstream bridge	Culture and institutions are shared externalized distinction infrastructures with verification costs.	Externalized symbols function semantically without interpreter, update rule, verification, or action-relevance channel.

2 Core ingredients from FDS

2.1 Active boundary relevance

The FDS core does not apply the deficit-dissipation-pruning cascade to every bounded object. It restricts the cascade to active-boundary systems. A minimal formal screen is

$$P\big(U(M_t, Y_t) \neq M_t\big) > 0, \qquad I(M_{t+1}; \ell_{t+k}) > 0 \tag{3}$$

for some $k > 0$. In empirical systems this relevance screen should be strengthened to an intervention or ablation test:

$$\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U)] \neq \mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U_\emptyset)], \tag{4}$$

where $U_\emptyset$ is a null, frozen, randomized, or identity update. This criterion is inherited from the FDS Core's active-boundary qualification [1]. In M0, it becomes the starting point for agency.

2.2 Capacity deficit and semantic demand

Let Ψ_{sem} be a pre-specified family of semantic task statistics: features of signals, contexts, actions, outcomes, and boundary states that must be preserved for a system to maintain a task within tolerance. Let

$$R_{\min}^{(\tau)}(\epsilon; \Psi_{\text{sem}}) \tag{5}$$

be the minimum number of bits per update window required to encode a sufficient semantic statistic to distortion tolerance ϵ. If C_{sem} is the maintained semantic memory capacity, define

$$\Delta_{\text{sem}}(t) = R_{\min}^{(\tau)}(\epsilon; \Psi_{\text{sem,t}}) - C_{\text{sem}}(t). \tag{6}$$

Positive semantic deficit does not prove that meaning fails. It says that without better compression, externalization, pruning, task relaxation, or resource expansion, actionable distinctions must be lost, merged, or degraded.

2.3 Notation alignment

Throughout this paper ℓ_{maint} denotes boundary-maintenance loss. It is not physical entropy production. The physical entropy ledger, when relevant, is denoted Σ_{phys} or $\dot{\Sigma}_{\text{phys}}$. Maintaining low ℓ_{maint} may require entropy or resource cost under physical bridge assumptions, but $\ell_{\text{maint}} \neq \Sigma_{\text{phys}}$. Likewise, q denotes a quotient map in the FDS Core; M0 uses $q_{\text{sem}} : \mathcal{D} \to T$ for a semantic quotient to avoid confusing semantic compression with topological or physical invariant quotients.

3 The agency-semantics dependency skeleton

Equation (2) is best understood as a constraint pyramid. Higher layers require stricter conditions on memory, verification, stability, and resource budgets. Under capacity deficit, they can degrade backward: meaning can collapse into attention, goal can collapse into reflex, value can collapse into salience, and culture can collapse into polluted external records.

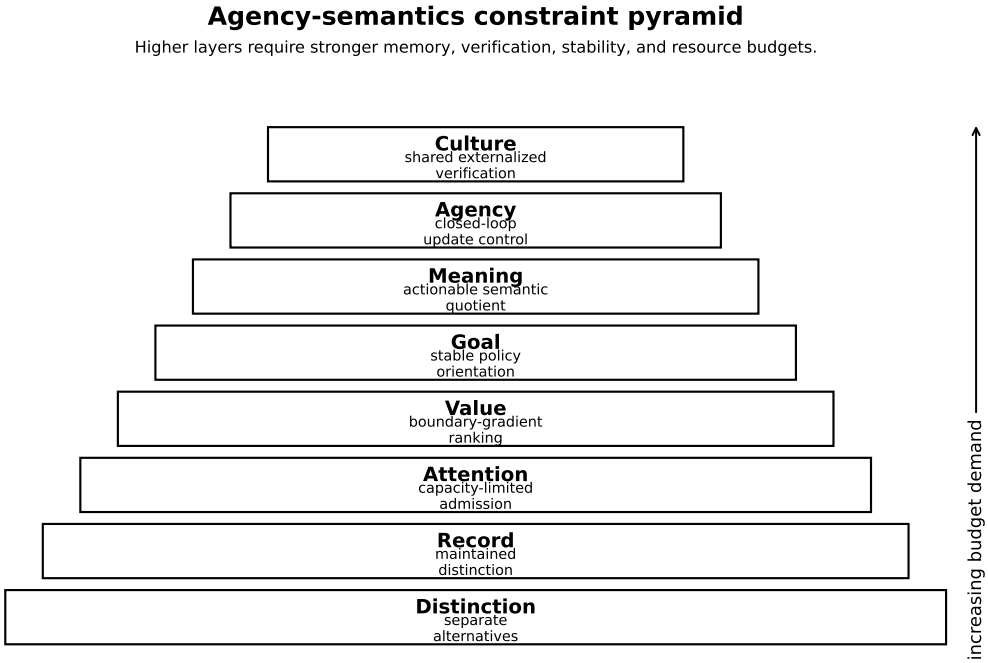

Figure 1: Synthetic normal-form illustration, not empirical evidence. The agency-semantics ladder is a dependency skeleton and constraint pyramid. Higher layers require stronger maintained records, verification capacity, stable policies, and externalized infrastructure.

Table 2: Agency-semantics constraint pyramid and failure modes under capacity deficit.

Layer	FDS operation	Main budget	Deficit failure mode
Attention	admission filtering	C_{att}	missed signal, distraction, tunnel vision
Value	boundary-gradient ranking	C_{eval}	misvaluation, risk blindness, reward hacking
Goal	stabilized policy orientation	M, τ	drift, reflex capture, unstable preference
Meaning	actionable semantic quotient	C_{sem}	semantic drift, hallucination, false compression
Agency	boundary-relevant update control	A, U, C_{verify}	proxy control, misalignment, collapse of autonomy
Culture	shared externalized distinctions	$E_{\text{ext}}, C_{\text{verify}}$	epistemic pollution, institutional rigidity

4 Attention as distinction admission

Let $\mathcal{D}_t = \{d_{t,1}, \ldots, d_{t,n}\}$ be a stream of candidate distinctions. Each distinction has a cost $c_t(d)$, including encoding, verification, and maintenance cost. An attention gate is a map

$$a_t : \mathcal{D}_t \to [0, 1], \tag{7}$$

subject to the expected capacity constraint

$$\sum_{d \in \mathcal{D}_t} a_t(d) c_t(d) \leq C_{\text{att}}(t). \tag{8}$$

Definition 1 (Attention)**.** Attention is capacity-limited admission of candidate distinctions into the update channel of an active finite system. It is not identical to salience: a salient distinction may fail admission if its verification or maintenance cost is too high, and a low-salience distinction may be admitted if it has high boundary relevance.

The attention allocation problem can be written as

$$a_t^* \in \arg\max_{a_t} \sum_{d \in \mathcal{D}_t} a_t(d) V_t^{\text{causal}}(d; k) \quad \text{s.t.} \quad \sum_d a_t(d) c_t(d) \leq C_{\text{att}}(t), \tag{9}$$

with $a_t(d) \in \{0, 1\}$ for hard admission or $a_t(d) \in [0, 1]$ for soft admission. Binary attention is a knapsack-like selection problem; soft attention is a fractional allocation problem.

Under high semantic deficit, the gate can become nonlinear and thresholded. One normal-form gate is

$$a_t(d) = \sigma\left(\beta_t[V_t(d; k) + \alpha n_t(d) - \gamma c_t(d) - \theta_t]\right), \qquad \beta_t = \beta_0 + \beta_\Delta[\Delta_{\text{sem}}(t)]_+, \tag{10}$$

where $n_t(d)$ is novelty and σ a logistic function. As Δ_{sem} grows, the gate steepens. This produces tunnel vision: only high-value or high-threat distinctions enter the update channel, while slower semantic structure is lost.

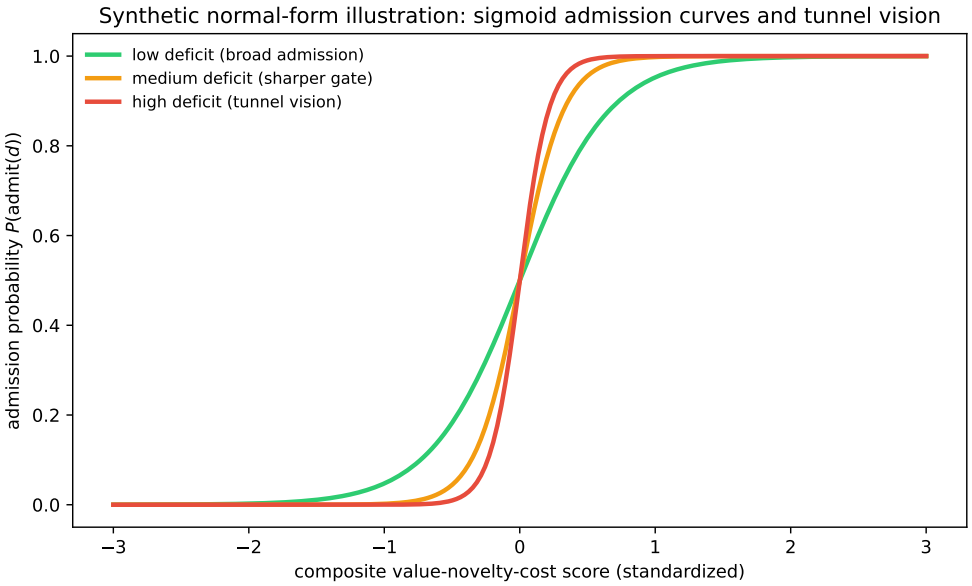

Figure 2: Synthetic normal-form illustration, not empirical evidence. Sigmoid admission curves under low, medium, and high semantic deficit. As Δ_{sem} increases, the gate steepens, producing tunnel vision: only high-value or high-threat distinctions enter the update channel while slower semantic structure is lost.

5 Value as causal boundary-gradient relevance

Value is not merely correlation with future loss. M0 distinguishes predictive value from causal value.

Definition 2 (Predictive value)**.** A distinction d has predictive value over horizon k if conditioning on it improves prediction of future boundary-maintenance loss:

$$V_t^{\text{pred}}(d; k) = \mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t] - \mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t, d] - \lambda_t c_t(d). \tag{11}$$

Definition 3 (Causal boundary-gradient value)**.** A distinction d has causal value if admitting it changes the expected future boundary-maintenance loss under the system's update and action channel:

$$V_t^{\text{causal}}(d;k) = \mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t, \text{do}(a_\emptyset)] - \mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t, \text{do}(\text{admit}(d))] - \lambda_t c_t(d). \tag{12}$$

The predictive version is an observational proxy. The causal version is the preferred FDS value notion, because a distinction that merely correlates with loss but cannot alter update, action, verification, or coordination is not yet valuable in the operational sense. Here $\text{do}(\text{admit}(d))$ denotes an intervention that forces $a_t(d) = 1$ through the admission gate and permits the downstream update and action policy to condition on the resulting maintained record. Similarly, $\text{do}(\neg\,\text{admit}(d))$ forces $a_t(d) = 0$ while leaving the rest of the gate, update, and action policy unchanged.

Near collapse thresholds, average loss is insufficient. Let ℓ_c be a critical boundary-loss threshold. Define a risk-weighted value

$$V_t^{\text{risk}}(d;k) = \Big(\mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t, \text{do}(\neg\,\text{admit}(d))] - \mathbb{E}[\ell_{\text{maint,t+k}} \mid M_t, \text{do}(\text{admit}(d))]\Big) + \alpha_t \Big(P(\ell_{\text{maint,t+k}} > \ell_c \mid M_t, \text{d} \tag{13}$$

where the first term is expected loss reduction and the second is collapse-risk reduction, both positive when admission reduces risk. This connects M0 to boundary-risk asymmetry: distinctions that move the system away from Phase-C collapse can dominate distinctions that only improve average performance. The critical tolerance ℓ_c is treated as fixed in M0. In adaptive systems it may itself be updated, negotiated, or misestimated; this is left to M2 and later governance applications.

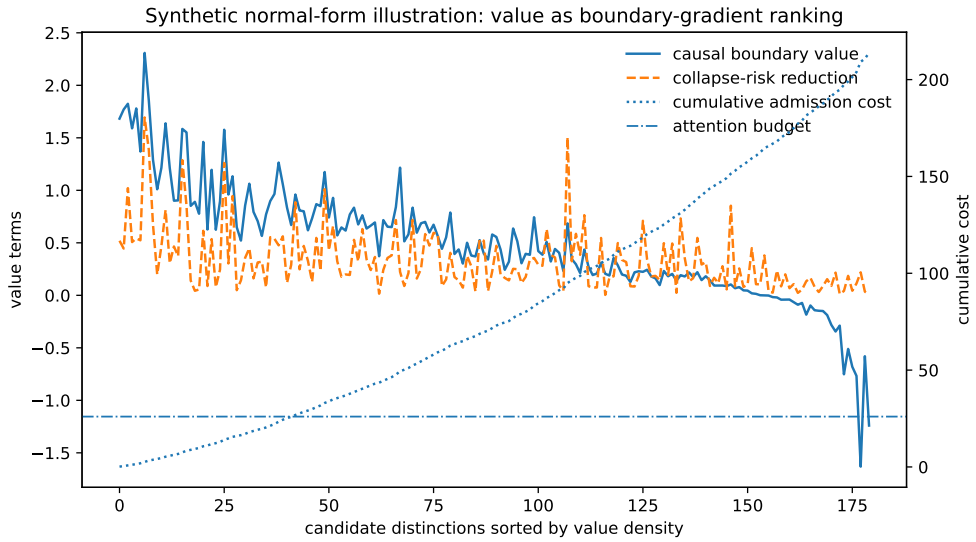

Figure 3: Synthetic normal-form illustration, not empirical evidence. Candidate distinctions are ranked by causal boundary-gradient value density, not by raw salience. Collapse-risk reduction can dominate average performance improvement near critical thresholds.

6 Goals as stabilized value rankings

A system can rank distinctions without having a goal. A reflexive controller may react moment by moment without maintaining an orientation across time. M0 treats goals as stabilized value rankings coupled to action policies.

Definition 4 (Goal)**.** A goal is a value ranking or policy orientation that remains sufficiently stable across update windows and perturbations to guide action selection over a specified horizon.

Goal stability is a register-time property: a goal must persist across finite update windows as a maintained ordering, rather than appearing only as an instantaneous reaction. This links M0 to the register-time framework of O2 [3].

Let $\pi_g(a \mid M_t)$ be the policy induced by a goal state and let Δ be a comparison horizon. A policy-based goal-stability index is

$$\mathrm{GSI}_\pi(t, \Delta) = \exp\Big(- D_{\mathrm{KL}}\big(\pi_g(\cdot \mid M_t) \| \pi_g(\cdot \mid M_{t+\Delta})\big)\Big), \tag{14}$$

so that $\mathrm{GSI}_\pi \in (0, 1]$ with $\mathrm{GSI}_\pi = 1$ for identical policies and $\mathrm{GSI}_\pi \to 0$ as the policy diverges. A rank-based version is

$$\mathrm{GSI}_r(t, \Delta) = 1 - D_{\mathrm{rank}}(\succ_t, \succ_{t+\Delta}), \tag{15}$$

where D_{rank} is a normalized rank distance in $[0, 1]$.

Proposition 1 (Reflex-preference-goal distinction)**.** *A one-step reaction may have predictive or causal value without being goal-like. A goal-like system must maintain a nontrivial stability index over an update horizon and couple that stability to action selection.*

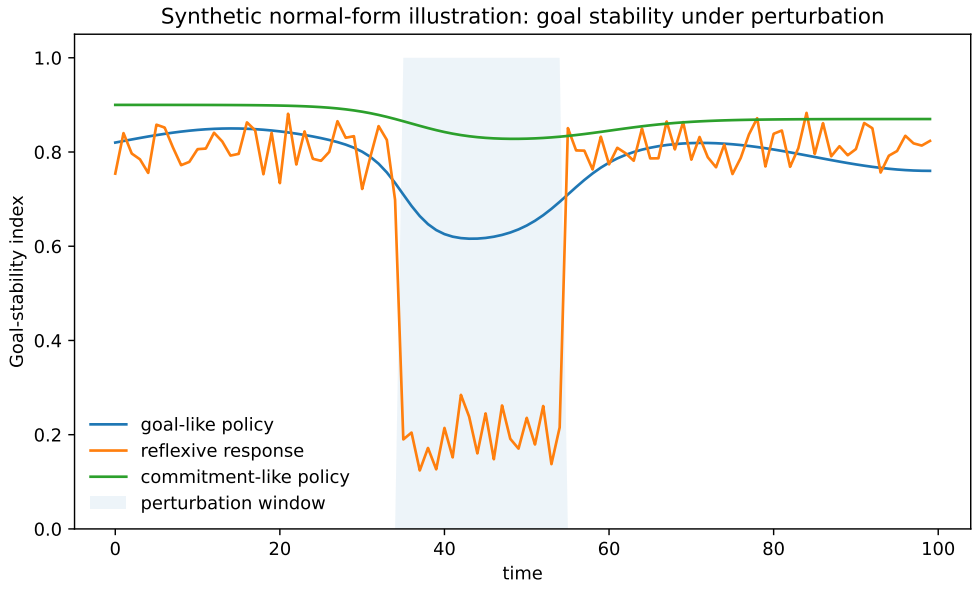

Figure 4: Synthetic normal-form illustration, not empirical evidence. A goal-like policy preserves a stable action orientation through perturbation better than a reflexive response. A commitment-like policy is even less sensitive but may become brittle if the environment changes.

7 Meaning as actionable semantic quotient

The phrase "actionable compressed distinction" is easily misunderstood if "actionable" is taken to mean immediate motor output. That is not the intended meaning. Actionable includes physical action, prediction, verification, coordination, compression of future search, preservation of an institutional boundary, or regulation of future update. A theorem, map, ritual, legal right, or social symbol may be actionable by constraining future inference and coordination, even when it does not trigger immediate motor behavior.

Definition 5 (Actionable compressed distinction)**.** A maintained representation m_d is an actionable compressed distinction for system S if it is a compressed record of a distinction d that preserves enough structure to change action, prediction, verification, coordination, or future update relative to a boundary-maintenance task.

Definition 6 (Policy-preserving semantic quotient)**.** Let $q_{\text{sem}} : \mathcal{D} \to T$ be a quotient map from candidate distinctions to semantic types. It is task-sufficient at tolerance ϵ for task family Ψ, horizon k, and context family $\mathcal{C}$ if there exists a quotient policy $\pi_q : T \times \mathcal{C} \to A$ such that for every $d \in \mathcal{D}$ and every admissible context $c \in \mathcal{C}$,

$$\mathbb{E}[\ell_{\text{maint,t+k}} \mid d, c, \text{do}(\pi_q(q_{\text{sem}}(d), c))] \leq \min_{a \in A} \mathbb{E}[\ell_{\text{maint,t+k}} \mid d, c, \text{do}(a)] + \epsilon. \tag{16}$$

This condition is stronger than requiring that the minimum achievable loss for two distinctions be close. It requires that a policy based on the quotient remain nearly optimal for each original distinction. This avoids the failure case where two distinctions have equal optimal loss but require different optimal actions.

Theorem 1 (Actionable meaning preservation)**.** *If q_{sem} is task-sufficient in the sense of Eq. (16), then replacing d by $q_{\text{sem}}(d)$ preserves actionable meaning up to ϵ for the specified task, horizon, context family, and action space.*

Proof. By the definition of a policy-preserving semantic quotient (Definition 6), for each original distinction d and admissible context c, the quotient policy $\pi_q(q_{\text{sem}}(d), c)$ achieves expected loss no more than ϵ above the best action available with access to d. Therefore the quotient retains all task-relevant action information needed to achieve near-optimal boundary-maintenance performance at the stated tolerance. Replacing d by $q_{\text{sem}}(d)$ therefore preserves actionable meaning up to ϵ. □

Remark 1 (Relation to abstraction and information bottlenecks). Equation (16) is close in spirit to state abstraction and bisimulation in reinforcement learning, and to task-relevant compression in the information bottleneck. M0 does not claim priority over those formalisms. It uses them as nearby mathematical idioms for a finite-boundary semantics bridge.

Example 1 (Red light)**.** A red traffic light compresses many visual details into a boundary-relevant affordance: stop, crossing risk, legal obligation, and coordination with others. The meaning is not the wavelength alone. It is the context-robust quotient that preserves action and coordination across drivers, pedestrians, vehicles, and legal settings.

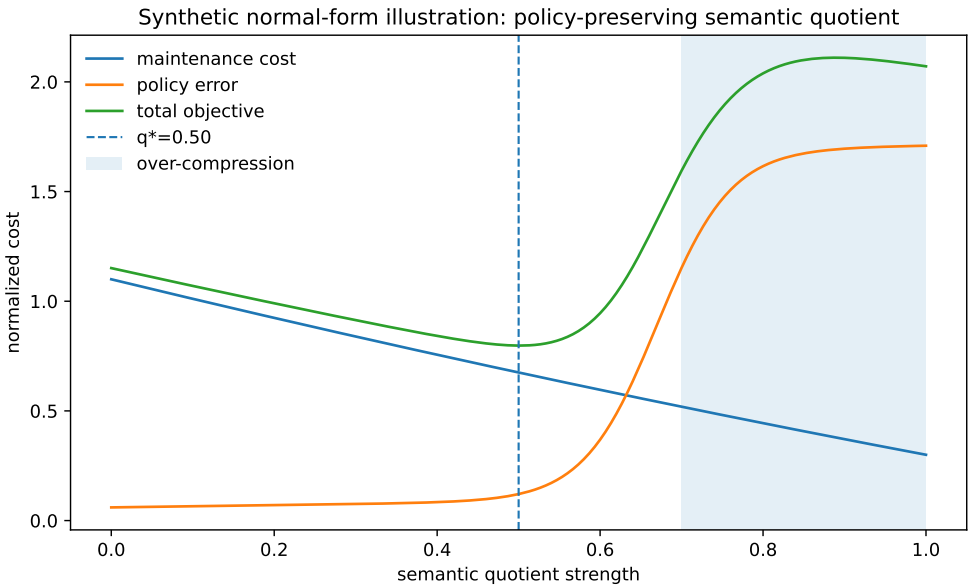

Figure 5: Synthetic normal-form illustration, not empirical evidence. Moderate semantic quotienting reduces maintenance cost while preserving policy performance. Over-compression destroys task-relevant distinctions and produces semantic drift or hallucination.

8 Semantic capacity deficit and collapse

A system can lose meaning without losing all signals. It may still attend to salient events while losing the goal, value, or policy context that makes them actionable. Define semantic capacity deficit as in Eq. (6). When $\Delta_{\text{sem}} > 0$, at least one exit must occur: better compression, pruning, externalization, task relaxation, verification outsourcing, or collapse into false compression.

Definition 7 (Semantic collapse)**.** Semantic collapse occurs when a system retains records or salience but loses the quotient structure required to preserve boundary-relevant action, prediction, verification, or coordination.

Examples include a stressed operator who notices alarms but loses system-level meaning, a long-context AI system that preserves local tokens while losing task sequence, and an institution that keeps labels while losing verification of what those labels do.

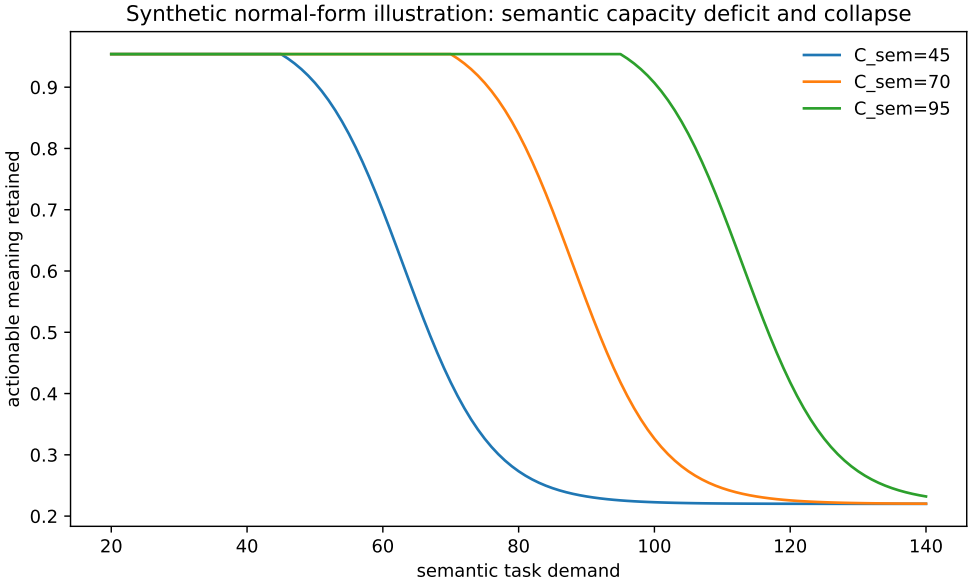

Figure 6: Synthetic normal-form illustration, not empirical evidence. As semantic task demand exceeds maintained semantic capacity, actionable meaning degrades. Signals can remain present while policy-preserving quotient structure collapses.

9 Agency as boundary-relevant update control

Definition 8 (Weak FDS agency)**.** A system has weak FDS agency if it selects actions from a finite action space in response to admitted distinctions.

Definition 9 (Strong FDS agency)**.** A system has strong FDS agency relative to boundary B, loss ℓ_{maint}, update rule U, and horizon k if there exists an admissible null update $U_\emptyset$ such that

$$\mathbb{E}[\ell_{\text{maint,t+k}} \mid \text{do}(U)] \neq \mathbb{E}[\ell_{\text{maint,t+k}} \mid \text{do}(U_\emptyset)]. \tag{17}$$

This is not a test of whether a system looks intelligent. It is a causal test of whether its updates matter for future boundary loss.

9.1 Boundary sensitivity

Agency classification is boundary-sensitive. A bare model, a model plus context, a model plus tools and memory, and a robot or institution embedded in an environment may receive different classifications. M0 therefore requires reports to state the accounting boundary.

Table 3: Boundary-sensitive agency classification.

Boundary choice	Possible classification	Attribution rule
Bare inference model	passive mapper or weak agency	Do not attribute strong agency if there is no durable boundary-relevant update.
Model + context window	weak coupled agency	Attribute behavior to the coupled prompt-context system.
Model + tools + writable memory	coupled active agency if updates affect future loss	Attribute agency to the coupled architecture, not base weights alone.
Robot / institution / environment loop	strong coupled agency if Eq. (17) holds	Report boundary, memory, update, verification, and loss mapping.

9.2 Verification load and self-verifying agency

Many systems appear agentic because an external host carries the verification load. Let Z_{outcome} be the outcome variable needed to determine whether actions reduce boundary loss. A simple verification deficit is

$$\Delta_{\text{verify}}(t) = R^{(\tau)}_{\text{min}}(\epsilon; \Psi_{\text{verify}}) - C^{\text{int}}_{\text{verify}}(t). \tag{18}$$

When $\Delta_{\text{verify}} > 0$, the system must externalize verification, rely on reward proxies, trust a host, relax the task, or operate blindly.

Definition 10 (Self-verifying agency)**.** A strong FDS agent is self-verifying at tolerance ϵ if it maintains an internal or coupled verification channel sufficient to estimate whether its actions reduce ℓ_{maint} within the task tolerance, without requiring an unmodeled host to perform the verification step.

Self-verification is not required for all strong agency. It is a stronger classification. This matters for artificial systems: if humans provide all outcome verification, the observed agency belongs to a coupled human-tool system.

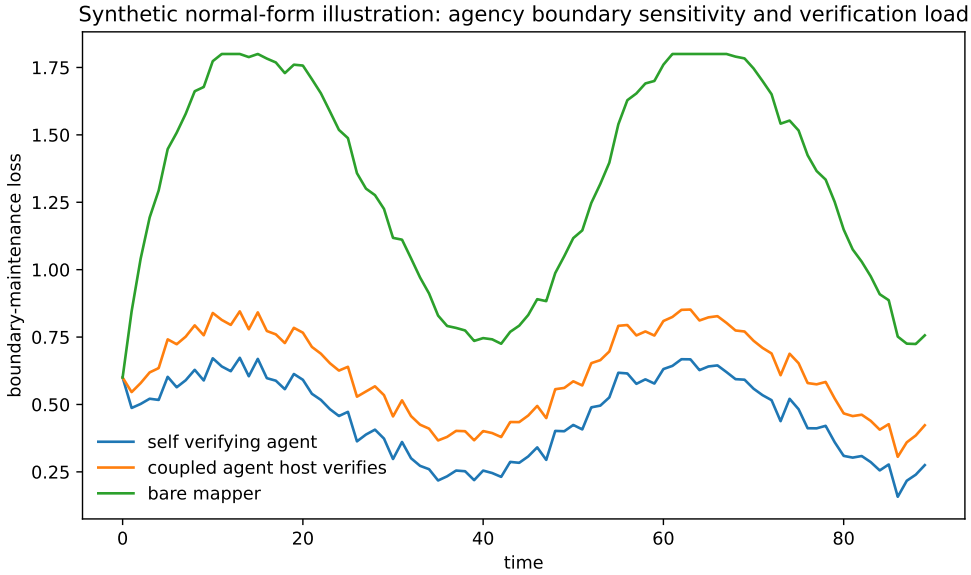

Figure 7: Synthetic normal-form illustration, not empirical evidence. Boundary-sensitive classification separates a self-verifying agent, a coupled agent whose host carries verification load, and a bare mapper that lacks durable boundary-relevant update.

10 Misalignment and epistemic pollution

10.1 Misalignment as action-effect divergence

Gradient dot products are coordinate-dependent and do not apply naturally to discrete action spaces. M0 therefore defines misalignment using normalized finite-difference effect vectors. For actions a_i relative to a baseline a_0, let

$$\Delta_H(a_i) = \mathbb{E}[\ell_H \mid \mathrm{do}(a_i)] - \mathbb{E}[\ell_H \mid \mathrm{do}(a_0)], \tag{19}$$

$$\Delta_D(a_i) = \mathbb{E}[\ell_D \mid \mathrm{do}(a_i)] - \mathbb{E}[\ell_D \mid \mathrm{do}(a_0)], \tag{20}$$

where H is the host and D the delegate. Define

$$\mathrm{Align}(H, D) = \frac{\langle \Delta_H, \Delta_D \rangle}{\|\Delta_H\| \, \|\Delta_D\|}. \tag{21}$$

For threshold $\eta > 0$, a delegated system is misaligned over the audited action set when

$$\mathrm{Align}(H, D) < -\eta. \tag{22}$$

This definition captures whether actions that help the delegate tend to harm the host, not merely whether internal objectives use different words.

Because Δ_H and Δ_D are defined as loss changes relative to baseline, beneficial actions have negative components. Alignment is positive when host and delegate loss effects point in the same direction, including the case in which both losses decrease.

10.2 Epistemic pollution as verification-bandwidth saturation

Externalized semantic environments can help finite systems by storing records outside internal memory. They can also pollute the shared distinction environment.

Definition 11 (Epistemic pollution)**.** Epistemic pollution is saturation of finite verification bandwidth by low-quality, adversarial, obsolete, redundant, or unverifiable distinctions in an externalized semantic environment.

Let $R^{(\tau)}_{\mathrm{verify}}(\epsilon; t)$ be the verification demand imposed by the external environment and $C^{\mathrm{avail}}_{\mathrm{verify}}(t)$ available verification capacity. Define

$$Z_{\mathrm{poll}}(t) = \frac{R^{(\tau)}_{\mathrm{verify}}(\epsilon; t)}{C^{\mathrm{avail}}_{\mathrm{verify}}(t)}. \tag{23}$$

When $Z_{\mathrm{poll}} > 1$, the system cannot verify all admitted distinctions. It must ignore, trust blindly, false-compress, outsource verification, or collapse into noise. False compression occurs when a quotient merges distinctions that should remain separate for the task, producing semantic drift, hallucination, category failure, or institutional rigidity.

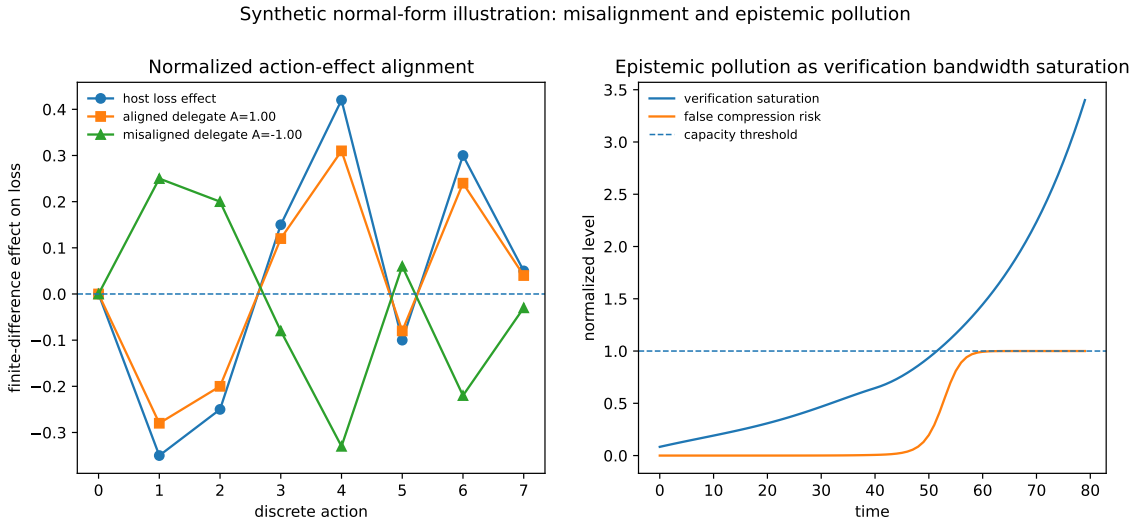

Figure 8: Synthetic normal-form illustration, not empirical evidence. Left: normalized finite-difference action effects distinguish aligned and misaligned delegated systems. Right: epistemic pollution appears when verification demand saturates finite verification bandwidth, forcing false compression.

11 Non-domain cases and attribution discipline

M0 is designed to prevent over-attribution. The following cases may process information or symbols without satisfying strong FDS agency or actionable meaning as defined above.

Table 4: Non-domain and limited-domain cases.

Case	Why it is not strong FDS agency or robust meaning by itself
Passive sensor	Records correlate with the world but have no boundary-relevant update control.
Static database	Stores distinctions, but action relevance appears only in a coupled user or institutional system.
Bare LLM inference call	Input-output mapping without durable self-updating boundary state or internal verification channel.
Random controller	Produces actions without expected boundary-loss reduction.
Lookup table	Mapping may be useful but lacks adaptive update, compression, or generalization unless embedded in a larger system.
Polluted archive	Preserves labels but not reliable verification, actionable quotient structure, or semantic maintenance.

12 Normal-form model and reproducibility

The accompanying code implements deterministic synthetic normal-form illustrations. The model is not empirical evidence. It is a consistency and visualization device that maps the definitions to simple state variables.

The model uses: a candidate distinction stream with novelty, threat, affordance, verification cost, and maintenance cost; a nonlinear attention gate whose steepness increases with semantic deficit; a causal boundary-gradient value score; a policy-preserving semantic quotient tradeoff; a goal-stability index under perturbation; boundary-sensitive agency rollouts; finite-difference misalignment; and epistemic pollution saturation. The random seed is fixed

in `code/generate_results.py`. CSV outputs are stored in `data/`; figure pairs are stored in `figures/`.

Table 5: Normal-form variable map. All entries are illustrative, not fitted empirical quantities.

Simulation variable	Paper definition	Interpretation
`causal_value`	Eq. (12)	expected boundary-loss reduction minus admission cost
`semantic_deficit`	Eq. (6)	task demand minus maintained semantic capacity
`admission_probability`	Eq. (10)	soft attention gate output
`policy_loss`	Eq. (16)	error induced by quotient compression
`goal_like_GSI`	Eq. (14)	stability of policy orientation under perturbation
`verification_load`	Eq. (18)	internal or external cost of action-outcome auditing
`saturation_ratio`	Eq. (23)	verification demand divided by verification capacity

13 Relation to existing fields

M0 is adjacent to several established fields.

Information theory and finite capacity. Shannon's information theory and rate-distortion theory define the fundamental limits of communication and compression under fidelity constraints [7, 8]. M0 uses these as formal boundaries for finite semantic capacity and semantic deficit.

Semiotics and pragmatics. Peircean and Morris-style semiotics analyze signs, objects, interpreters, and action consequences [19, 20]. M0 does not replace semiotics. It specifies a finite-system bridge: a sign functions as a maintained quotient only relative to an interpreter boundary, update rule, action space, and verification channel.

Information bottleneck and semantic compression. The information bottleneck formalizes compression that preserves target-relevant information [9]. M0's semantic quotient is similar but explicitly tied to action, policy preservation, and boundary-maintenance loss.

Rational inattention and bounded rationality. Rational inattention treats information acquisition as costly [10, 11]. M0 generalizes this into distinction admission: attention is not only costly information acquisition, but capacity-limited update admission under maintenance constraints.

Reinforcement learning, state abstraction, and bisimulation. State abstraction and bisimulation preserve decision-relevant structure in Markov decision processes [13–15]. M0's Theorem 1 is closest to this family: meaning-preserving quotients must preserve near-optimal action, not merely compress observations.

Causal representation and intervention. Pearl-style causal models emphasize interventions rather than correlation [12]. M0 uses this distinction for causal value and agency: a distinction is operationally valuable only when admission or action changes expected boundary loss.

Affordances, empowerment, active sensing, and active inference. Affordance theory stresses action possibilities [16]; empowerment measures potential influence over future states [17]; active inference and predictive processing emphasize action-perception loops under uncertainty [18]. M0 is compatible with these perspectives but centers finite distinction capacity, verification load, and boundary-maintenance loss.

Game, trust, and verification. Earlier unpublished distinction-budget game drafts treated cooperation, parasitism, and trust in terms of monitoring costs. M0 uses the more conservative FDS language: trust and delegation reduce repeated verification cost only when maintained records remain reliable and when externalized verification does not saturate capacity.

Recent interfaces. Recent work sharpens several concepts used in M0. Causal bisimulation and causal state abstraction study how task-relevant structure can be compressed while preserving decision performance [21, 22]. Goal-oriented semantic communication similarly treats semantics as task-relevant compression under bandwidth, latency, and power constraints [25]. Recent work on active inference distinguishes variational free energy from thermodynamic free energy, clarifying the physical cost of active agents [23]. Recent cognitive-economic work connects machine learning, costly learning, and rational inattention [24]. Finally, LLM-agent surveys provide a practical background for the boundary-sensitive agency distinction used here: agency may belong to a coupled system of model, tools, memory, user, and environment rather than to a bare model call [26].

14 Protocols and tests

Protocol 1 (Attention admission audit)**.** Pre-register candidate distinctions, costs, and boundary-relevance variables. Vary load or capacity and test whether admitted distinctions track causal boundary value better than raw salience.

Protocol 2 (Value intervention audit)**.** Compare predictive and causal value. A distinction with high predictive value but no effect under admission or action should be demoted to observational relevance.

Protocol 3 (Goal stability audit)**.** Estimate GSI_π or GSI_r under perturbations. Reflexive systems may react strongly while showing low stability across update windows.

Protocol 4 (Semantic quotient audit)**.** Test whether a compressed representation admits a quotient policy satisfying Eq. (16). Over-compression should increase task loss or semantic drift.

Protocol 5 (Agency intervention audit)**.** Compare $\mathrm{do}(U)$ against $\mathrm{do}(U_\emptyset)$ under a stated boundary. If only the extended architecture passes, classify agency as coupled rather than intrinsic to the narrow component.

Protocol 6 (Misalignment and verification audit)**.** Compute finite-difference action-effect alignment for host and delegate losses. Separately audit verification load to identify systems whose apparent agency is supplied by an external host.

15 Limitations and falsification

M0 is intentionally limited. It does not establish empirical theories of attention, value, goal, or meaning by definition. It provides mappings that must survive operational audit. The framework is weakened or demoted under any of the following results:

1. robust attention-like behavior under a specified mapping shows no capacity-limited admission or selection;
2. value-like ranking has no relationship to future boundary loss, task success, risk reduction, or resource relevance;
3. goal-like behavior persists without memory, ranking stability, policy orientation, or action direction over time;
4. compressed representations guide no action, prediction, verification, coordination, or boundary maintenance;
5. strong agency is attributed to systems whose updates have no causal effect on future boundary-maintenance loss;
6. delegated systems remain aligned despite persistent negative action-effect alignment and verification failure;
7. externalized semantic environments impose no retrieval, verification, maintenance, or pollution costs.

16 Conclusion

FDS-M0 defines the agency-semantics spine of Distinction Theory. Its central chain is:

$$\text{finite distinction} \to \text{admission} \to \text{boundary-gradient ranking} \to \text{stable policy} \to \text{actionable semantic quotient} \tag{24}$$

The chain is not a claim that semantics has been solved or reduced to physics. It is an operational interface. Attention, value, goal, meaning, and agency are treated as roles played by distinctions in active finite systems that must maintain boundaries under capacity and resource constraints.

This spine prepares later M-series and civilization-layer papers. M1 can focus on attention as distinction admission; M2 on value and goal as boundary-gradient ranking; M3 on meaning as actionable semantic quotient; M5 on trust as delegated verification; A2 on AI alignment as boundary-compatible externalized agency; S2 on epistemic pollution; and G3/G4 on science and civilization memory as large-scale verification infrastructures.

Remark 2 (Cultural invariant candidate). A cultural invariant is a shared externalized semantic quotient that remains usable across agents, contexts, and update windows for coordination or institutional boundary maintenance. M0 only defines the interface; S/G-series papers develop the civilization-scale theory.

The central discipline remains unchanged: every domain claim must state its boundary, memory, update rule, capacity, loss, verification channel, and failure condition.

Code and data availability

The deterministic synthetic normal-form code, figures, and CSV outputs are included in the replication package. Run `python code/generate_results.py` from the paper directory to regenerate all figures and data.

AI assistance disclosure

The author used AI assistance for drafting, editing, simulation scaffolding, and consistency checks. The author reviewed and selected the final claims, definitions, equations, and interpretations.

A M-series dependency map

Table 6: How M0 seeds the later M-series.

Future paper	Interface supplied by M0
M1 Attention	distinction admission, attention allocation, tunnel vision
M2 Value/Goal	causal boundary-gradient value, risk asymmetry, goal-stability index
M3 Meaning	task-sufficient semantic quotient, over-compression, semantic collapse
M4 Learning	distinction refinement and compression improvement under novelty
M5 Trust	delegated verification, monitoring cost, externalized confidence records
M6 Power	control over admission channels, verification infrastructure, and boundary updates
M7 Culture	shared externalized distinction infrastructure
M8 Beauty	high-compression distinction with low maintenance and high transfer
M9 Education	intergenerational distinction transmission and verification
M10 Law/Rights	boundary protocols and protected distinction claims

B Notation summary

Symbol	Meaning
$\mathcal{D}_t$	candidate distinction stream
$a_t(d)$	attention/admission weight for distinction d
$c_t(d)$	encoding, verification, and maintenance cost of d
C_{att}	attention admission capacity
V_t^{pred}	predictive value, observational relevance to future loss
V_t^{causal}	causal value, effect of admitting d on future loss
V_t^{risk}	risk-weighted value near collapse thresholds
GSI	goal-stability index
q_{sem}	semantic quotient preserving action-relevant structure
Δ_{sem}	semantic capacity deficit
ℓ_{maint}	boundary-maintenance loss
Σ_{phys}	physical entropy-production ledger, distinct from ℓ_{maint}
Δ_{verify}	verification deficit
Z_{poll}	epistemic pollution / verification saturation ratio

References

[1] Y. Wu, *Active Finite Distinction Systems: A Formal Core for Boundary Maintenance under Finite Capacity*, Zenodo (2026), doi:10.5281/zenodo.20158923.

[2] Y. Wu, *Observer as a Finite Distinction Register*, Zenodo (2026), doi:10.5281/zenodo.20248792.

[3] Y. Wu, *Time as Irreversible Distinction Update*, Zenodo (2026), doi:10.5281/zenodo.20249369.

[4] Y. Wu, *Capacity Overflow and Effective Stochasticity*, Zenodo (2026), doi:10.5281/zenodo.20250367.

[5] Y. Wu, *Boundary-Maintaining Self-Organizing Systems under Finite Capacity*, Zenodo (2026), doi:10.5281/zenodo.20253151.

[6] Y. Wu, *Boundary Maintenance and the Second Law under Finite Memory*, Zenodo (2026), doi:10.5281/zenodo.20255129.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal* 27, 379–423 and 623–656 (1948).

[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley, 2006).

[9] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing* (1999).

[10] H. A. Simon, "A behavioral model of rational choice," *Quarterly Journal of Economics* 69, 99–118 (1955).

[11] C. A. Sims, "Implications of rational inattention," *Journal of Monetary Economics* 50, 665–690 (2003).

[12] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (Cambridge University Press, 2009).

[13] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (MIT Press, 2018).

[14] L. Li, T. J. Walsh, and M. L. Littman, "Towards a unified theory of state abstraction for MDPs," in *Proceedings of ISAIM* (2006).

[15] N. Ferns, P. Panangaden, and D. Precup, "Metrics for finite Markov decision processes," in *Proceedings of UAI* (2004).

[16] J. J. Gibson, *The Ecological Approach to Visual Perception* (Houghton Mifflin, 1979).

[17] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Empowerment: A universal agent-centric measure of control," in *IEEE Congress on Evolutionary Computation* (2005).

[18] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience* 11, 127–138 (2010).

[19] C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, Vols. 1–2, edited by C. Hartshorne and P. Weiss (Harvard University Press, 1931).

[20] C. W. Morris, *Foundations of the Theory of Signs* (University of Chicago Press, 1938).

[21] Z. Wang, C. Wang, X. Xiao, Y. Zhu, and P. Stone, "Building Minimal and Reusable Causal State Abstractions for Reinforcement Learning," *Proceedings of the AAAI Conference on Artificial Intelligence* **38**, 15778–15786 (2024), doi:10.1609/aaai.v38i14.29507.

[22] X. Li, S.-O. Kaba, and S. Ravanbakhsh, "On the Identifiability of Causal Abstractions," arXiv:2503.10834 (2025), doi:10.48550/arXiv.2503.10834.

[23] C. Fields, A. Goldstein, and L. Sandved-Smith, "Making the Thermodynamic Cost of Active Inference Explicit," *Entropy* **26**, 622 (2024), doi:10.3390/e26080622.

[24] A. Caplin, D. Martin, and P. Marx, "Modeling Machine Learning: A Cognitive Economic Approach," *Journal of Economic Theory* **224**, 105970 (2025), doi:10.1016/j.jet.2025.105970.

[25] T. M. Getu, G. Kaddoum, and M. Bennis, "A Survey on Goal-Oriented Semantic Communication: Techniques, Challenges, and Future Directions," *IEEE Access* **12**, 51223–51274 (2024), doi:10.1109/ACCESS.2024.3381967.

[26] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, "A Survey on Large Language Model Based Autonomous Agents," *Frontiers of Computer Science* **18**, 186345 (2024), doi:10.1007/s11704-024-40231-1.