Active Finite Distinction Systems as a Criterion for Artificial Agency

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

Large predictive models can transform inputs into high-quality outputs without necessarily becoming agents. This paper proposes an operational criterion for artificial agency based on *active finite distinction systems* (FDS): finite-capacity systems that maintain a boundary through state-dependent updates under resource constraints. The framework distinguishes passive mapping from artificial agency by requiring active boundary maintenance, durable update participation, causal loop closure, and capacity-deficit management through pruning, externalization, task relaxation, or compression. The central technical object is an FDS tuple

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

where B is an operational boundary, M is internal memory or model state, U is an update operator, ℓ is boundary-maintenance loss, Φ is a resource budget, $\mathcal{P}$ is a perturbation/pruning family, and τ is an update timescale. Capacity deficit is formulated as a task-relevant rate-distortion gap $\Delta_\varepsilon(\tau) = R^{(\tau)}_{\min}(\varepsilon) - C_S$, rather than as an informal mismatch between a model and the world. Physical dissipation is treated only through an explicit Landauer bridge: logically irreversible updates incur an informational heat floor proportional to $H(M_t \mid M_{t+1}, Y_t)$ under standard thermodynamic conditions. We derive an agency taxonomy separating passive mappers, scaffolded tool systems, adaptive controllers, and strong FDS-agents. We also give an operational test protocol using update ablation, action-to-future-state influence, capacity-deficit estimation, pruning/externalization efficiency, and resource-governed persistence. A minimal gridworld MDP illustrates how a one-bit capacity deficit induces an irreducible action-error lower bound, and a benchmark protocol is proposed for testing FDS-agency under bounded resources. The result is not a claim that scaling is irrelevant or that any named architecture cannot become agentic. It is a formal criterion for when a system containing predictive components qualifies as an artificial agent rather than merely a passive mapper.

Keywords: artificial agency; active finite systems; boundary maintenance; rate-distortion; transfer entropy; active pruning; external memory; resource-bounded agents; LLM agents; active inference.

INTRODUCTION

The passive-mapper problem

A central question in artificial intelligence is no longer whether a machine can produce intelligent-looking outputs. Large language models, world models, simulators, planners, and tool-using systems already perform tasks that previously appeared to require flexible reasoning. The harder question is structural: *when does a predictive or generative system become an agent rather than a passive mapper?*

A passive mapper transforms inputs into outputs. It may be highly competent, statistically calibrated, and useful. Yet input-output competence alone does not imply that the system maintains a boundary, updates durable self-relevant state, regulates its resources, prunes accumulated complexity, or causally participates in the future conditions it must survive. A model that can answer questions about a robot's battery is not thereby a system that monitors, protects, allocates, or recharges that battery. The difference is not prediction accuracy; it is the presence or absence of a maintenance loop connecting internal update, action, resource constraint, and future viability.

This paper develops an AI-facing formalization of that distinction. The proposal is that artificial agency requires more than prediction, generation, or optimization over an externally supplied objective. It requires that a finite system maintain an operational boundary under limited representational capacity and limited resources. This leads to a class of systems called *active finite distinction systems* (FDS). An FDS is not defined by being biological, conscious, embodied, or made from any particular substrate. It is defined by the structure of the maintenance problem it faces.

The FDS entry point

The formal core used here is the active finite distinction system

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

where X is internal state, E is environmental state, B is a boundary or interface variable, M is memory or model state, Y is the observation channel, A is the action or

boundary-update space, U is the internal update operator, π is a finite distinction projection or coarse-graining, ℓ is boundary-maintenance loss, Φ is a resource budget, $\mathcal{P}$ is an admissible family of pruning, perturbation, or degradation operators, and τ is the update timescale.

The tuple is deliberately operational. In AI systems, B may be a robot body, a software service boundary, a memory and tool boundary, an API-level identity, a security perimeter, or a task-level viability envelope. M may include parameters, context state, working memory, persistent memory, maps, caches, learned skills, policies, or externalized records. A may include physical actions, API calls, tool invocations, memory writes, model-update operations, self-repair operations, or resource-allocation decisions. The central question is whether these components form an active maintenance loop rather than a passive input-output channel.

Contributions

This paper makes seven contributions.

1. It reformulates artificial agency in terms of active finite distinction systems, making boundary maintenance and resource-bounded persistence central rather than optional.
2. It defines an active-boundary criterion using both a formal relevance screen and an empirical intervention/ablation strengthening.
3. It replaces informal claims about environmental complexity with a task-relevant rate-distortion capacity deficit.
4. It distinguishes passive mappers, scaffolded systems, adaptive controllers, and strong FDS-agents.
5. It derives a passive-mapper non-qualification proposition and a prune–externalize–collapse trichotomy for AI architectures under persistent capacity deficit.
6. It gives an operational test protocol for evaluating artificial agency using update ablation, action-to-future-state influence, capacity-deficit estimation, pruning/externalization efficiency, and resource-governed persistence.
7. It proposes benchmark desiderata and metrics for deep causal tracking tasks under bounded resources, including viability, resource efficiency, pruning gain, externalization gain, and ablation gap.

What is not claimed

The paper does not claim that current large language models have no intelligence. It does not claim that scaling is useless. It does not claim that artificial general intelligence is impossible. It does not claim that a system containing a fixed predictive model is necessarily passive; a fixed model embedded inside a tool-using, memory-writing, self-monitoring, resource-governed architecture may be part of an active system. It does not claim to solve consciousness. It does not derive physical law from the distinction primitive. The claims below are conditional, operational, and targeted at the artificial-agency question.

RELATED WORK

Agent theory and embodied AI

Classical accounts of agency distinguish agents from passive computational processes by emphasizing autonomy, situatedness, reactivity, proactivity, and social ability [7]. Dynamical systems approaches treat agency as a closed agent-environment loop rather than as isolated computation [8]. Embodied AI emphasizes that intelligence is shaped by sensorimotor coupling and environmental feedback rather than only by internal representation [9]. More recent formal accounts emphasize individuality, normativity, asymmetry, and spatio-temporal organization in defining agency [10].

The FDS approach is compatible with these traditions but isolates a different axis: the cost of maintaining a boundary under finite capacity. A system may be reactive or embedded without being strongly self-maintaining; conversely, a software architecture may lack a biological body but still maintain an operational boundary through memory, tools, monitoring, repair, and resource governance.

Active inference, Markov blankets, and autonomy

The free-energy principle and active inference model self-organizing systems as minimizing variational free energy through perception and action [11, 12]. Markov blankets formalize statistical boundaries separating internal and external states [13]. These frameworks are central to contemporary discussions of autonomy and agent-environment coupling.

The FDS framework differs in emphasis. It does not begin with belief updating or variational inference. It begins with active boundary maintenance under finite representational capacity. The FDS chain asks what follows when the task-relevant rate-distortion demand exceeds internal capacity: approximation, residual error, complexity pressure, irreversible update costs under physical implementation, and eventual pruning, externalization, task relaxation, or collapse. Active inference can supply one implementation of FDS-like dynamics, but FDS

is not restricted to Bayesian agents or variational objectives.

Control, empowerment, and transfer entropy

Control-theoretic and information-theoretic measures of agency often ask whether an action channel has causal influence over future states. Empowerment measures the channel capacity from actions to future sensor states [14]. Transfer entropy measures directed information transfer from one process to another [15]. In this paper, action-to-future-state influence is used as a marker of causal loop closure, but not as a sufficient condition for full agency. A thermostat may causally affect future temperature, yet it lacks the richer capacity-deficit management, active pruning, and resource-governed persistence required for strong FDS-agency.

LLM agents and tool-using systems

Recent LLM-based agents combine language models with planning, external tools, memory, environments, and feedback loops [16–19, 46]. Benchmarks show that current models perform well on single-turn tool calls but degrade under memory demands and long-horizon tasks [46]. Recent work on dynamic code generation as action further demonstrates externalization: agents generate Python functions at runtime rather than selecting from a fixed action set [47]. These systems illustrate why the base-model question and the coupled-system question must be separated. A feed-forward model evaluated at inference time may be passive; the larger architecture containing it may not be. The FDS criterion is explicitly system-level: the unit of agency is the coupled architecture that maintains boundaries, updates durable state, acts causally, manages resources, and handles accumulated complexity.

Information theory, rate-distortion, and thermodynamics

Rate-distortion theory studies the minimal coding rate needed to represent a source under a distortion constraint [22–24]. Landauer's principle links logically irreversible erasure to a thermodynamic heat cost under specified physical conditions [25, 26]. Stochastic thermodynamics generalizes entropy production and information thermodynamics in nonequilibrium systems [27–29]. The FDS formalism uses rate-distortion to define task-relevant capacity deficit and uses Landauer only as a physical bridge for irreversible updates, not as a claim that every act of inference directly dissipates $k_B T \ln 2$ per represented bit.

Autonomous agents and rational agency

Classical AI and agent theory define autonomous agents as systems that perceive their environment, make decisions, and act to achieve goals [30, 31]. Reinforcement-learning formalisms treat agents as policy-optimizing systems that maximize expected cumulative reward [32]. These traditions focus on decision-making under uncertainty and goal-directed behavior. The FDS criterion shifts emphasis from goal optimization to boundary maintenance: the central question is not whether a system maximizes a reward function, but whether it maintains the conditions of its own persistence under finite capacity. A reward-maximizing system that does not update or repair its own decision boundary is a passive mapper with respect to its own viability.

Memory-augmented and continual-learning systems

Neural network architectures augmented with external memory, attention, and retrieval mechanisms extend the effective capacity of static models [33–35]. Continual-learning and lifelong-learning systems address the problem of accumulating knowledge without catastrophic forgetting [36–38]. These lines of work are directly relevant to the FDS framework: durable memory and active consolidation are components of capacity-deficit management, and catastrophic forgetting is a special case of maintenance failure under representational pressure. The FDS criterion adds the requirement that memory and consolidation decisions be causally connected to boundary-maintenance loss, not merely to task accuracy.

Resource-rationality and bounded rationality

Bounded rationality recognizes that agents operate under cognitive, computational, and informational constraints [39, 40]. Resource-rational analysis models agents as approximately optimal under resource limitations [41, 42]. These perspectives are close to the FDS approach but focus on decision optimality under constraints rather than on boundary maintenance. The FDS framework asks what structures a system must have to persist while managing capacity deficit, rather than how close its decisions come to a resource-rational optimum.

AI safety, autonomy, and self-maintenance

Discussions of AI safety increasingly distinguish between capability, alignment, learned optimization, and systems that model their own role in deployment contexts [43–45]. These concerns overlap with the FDS criterion: a system that does not monitor or maintain its

own operational boundary cannot be expected to recognize changes in its own function, resource state, or goal integrity. The FDS framework does not solve the alignment problem, but it provides structural conditions for the kind of self-maintenance that safety properties may presuppose.

ACTIVE FINITE DISTINCTION SYSTEMS

The system tuple

Definition 1 (Active finite distinction system)**.** *An active finite distinction system is a tuple*

$$S = (X, E, B, M, Y, A, U, \pi, \ell, \Phi, \mathcal{P}, \tau), \tag{2}$$

where:

- X *is the internal state space of the system;*
- E *is the environmental state space;*
- B *is the boundary or interface variable separating system from environment;*
- M *is the internal memory, model, or representational state space;*
- Y *is the observation channel available at the boundary;*
- A *is the action or boundary-update space;*
- U *is the update map, typically* $U : M \times Y \to M$ *or a stochastic kernel* $P(M_{t+1} \mid M_t, Y_t)$*;*
- π *is the finite distinction projection, coarse-graining, or partition map;*
- ℓ *is the boundary-maintenance loss function;*
- Φ *is the finite resource or free-energy budget functional;*
- $\mathcal{P}$ *is the admissible family of pruning, coarse-graining, perturbation, or degradation operators;*
- τ *is the update timescale converting per-step information loss into rates or powers.*

Remark 1 (Why τ matters)**.** *Without a timescale, information-theoretic quantities remain dimensionless counts. With* τ*, per-update erasure and capacity demand can be expressed as rates and connected to resource or power budgets. This is essential for distinguishing a static representation from a maintained active system.*

AI interpretation of the tuple

For artificial systems, the tuple can be read as follows. X includes all internal operational state relevant to future behavior. M includes learned parameters, context, scratchpads, persistent memory, maps, policies, cached tool outputs, and externalized records treated as part of the agent architecture. Y includes sensors, prompts, API responses, environment observations, telemetry, or monitoring signals. A includes actions, tool calls, memory writes, environment modifications, compute-allocation decisions, and self-maintenance operations. ℓ measures boundary-maintenance or viability loss: task failure, resource exhaustion, context corruption, actuator failure, memory inconsistency, policy degradation, security breach, or loss of operational continuity. Φ may be energy, money, compute budget, time, memory, latency, user trust, or any resource without which the system cannot maintain the relevant boundary.

The tuple does not require that all parts live inside one neural network. A model, memory store, retriever, planner, robot, monitoring service, database, and tool environment can jointly form a single FDS if the coupled architecture maintains an operational boundary through durable state-dependent updates.

ACTIVE BOUNDARIES AND CAUSAL QUALIFICATION

Formal relevance condition

Definition 2 (Active boundary)**.** *An FDS has an active boundary if its update rule is nontrivial and is relevant to future boundary-maintenance loss. Formally,*

$$\mathbb{P}\big(U(M_t, Y_t) \neq M_t\big) > 0, \tag{3}$$

and there exists $k > 0$ *such that*

$$\mathrm{I}(M_{t+1}; \ell_{t+k}) > 0. \tag{4}$$

The first condition excludes systems with no update channel. The second condition excludes irrelevant internal noise. A random bit flip inside a machine is not active boundary maintenance unless it is statistically related to future boundary loss.

Intervention strengthening

Mutual information is a relevance screen, not a causal proof. A passive thermometer may encode future loss without participating in loss reduction. Therefore empirical AI applications require an intervention or ablation strengthening.

Criterion 1 (Causal active-boundary qualification)**.** *For empirical artificial systems, an update channel U participates in active boundary maintenance only if there exists an admissible null update channel $U_\emptyset$, such as freezing, randomizing, masking, or identity-updating the relevant state, and some $k > 0$ such that*

$$\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U)] \neq \mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U_\emptyset)]. \tag{5}$$

Remark 2 (Update ablation in AI)**.** *In an AI system, $U_\emptyset$ may disable memory writes, freeze tool feedback, randomize planner state, remove self-monitoring, mask resource telemetry, or prevent policy updates. If future boundary-maintenance loss is unchanged under such ablation, the removed update channel was not part of active agency for the tested task class.*

Passive boundaries

Definition 3 (Passive boundary)**.** *A system has a passive boundary when its persistence is supported by static constraints or stable configurations without task-directed internal updating. Passive persistence may be real and stable, but it does not enter the deficit–dissipation–maintenance cascade unless an active update channel is introduced.*

A read-only database, an inert file, a static model checkpoint, or a frozen feed-forward mapping has a boundary in some descriptive sense, but it does not thereby qualify as an active FDS-agent. It may become part of an active system when coupled to update, monitoring, action, and resource-governance loops.

CAPACITY DEFICIT AS A RATE-DISTORTION GAP

Representational capacity

Definition 4 (Internal representational capacity)**.** *For a finite memory or model state space M, define the internal representational capacity*

$$C_S = \log_2 |M|. \tag{6}$$

When M is continuous or effectively continuous, C_S is replaced by the appropriate operational capacity under finite resolution, noise, coding, quantization, compression, or resource constraints.

Proposition 1 (Finite distinction representation)**.** *For any FDS with finite internal capacity C_S,*

$$\mathrm{I}(E_t; M_t) \leq C_S. \tag{7}$$

Proof. By the data-processing inequality and the entropy bound on a finite memory state,

$$\mathrm{I}(E_t; M_t) \leq H(M_t) \leq \log_2 |M| = C_S.$$

□

This proposition does not say that the environment is finite. It says that the maintained internal distinction structure is finite.

Admissible task statistics

Definition 5 (Admissible task statistics)**.** *Let Ψ be a pre-registered family of admissible statistics $\psi : (E, B) \to Z$ causally accessible through the observation channel Y. A statistic $Z_t = \psi(E_t, B_t)$ is task-relevant if it captures information needed to maintain the boundary or task within a stated loss tolerance. Trivial constant maps are excluded unless the task itself is trivial. The family Ψ is fixed by the task specification and sensor-action architecture, not chosen after observing failure.*

For AI systems, typical admissible families include:

1. **Control/homeostasis tasks:** statistics sufficient to select actions that keep a boundary variable B_t inside a viable set $\mathcal{V}$ up to tolerance ε.
2. **Predictive tracking tasks:** statistics sufficient to predict a pre-specified future observable Y_{t+k} or target variable Z_{t+k} under a fixed loss function.
3. **Information-bottleneck tasks:** compressed statistics preserving task-relevant information about a pre-specified target Z under a stated rate-distortion or information-bottleneck objective.

Capacity deficit

Definition 6 (Rate-distortion demand)**.** *For an admissible statistic $\psi \in \Psi$, update window τ, and tolerated distortion ε, let*

$$R^{(\tau)}_{\psi(E,B)}(\varepsilon) \tag{8}$$

be the minimum number of bits per update window required to encode $\psi(E, B)$ so that expected boundary-maintenance distortion remains below ε. Define

$$R^{(\tau)}_{\min}(\varepsilon) = \inf_{\psi \in \Psi} R^{(\tau)}_{\psi(E,B)}(\varepsilon). \tag{9}$$

Definition 7 (Capacity deficit)**.** *The capacity deficit of an FDS at tolerance ε over update window τ is*

$$\Delta_\varepsilon(\tau) = R^{(\tau)}_{\min}(\varepsilon) - C_S. \tag{10}$$

If $\Delta_\varepsilon(\tau) > 0$, the system is in task-relevant capacity deficit at tolerance ε.

Theorem 1 (Capacity deficit theorem)**.** *If*

$$R^{(\tau)}_{\min}(\varepsilon) > C_S, \tag{11}$$

then no purely internal model M_t of capacity C_S can encode even the least demanding admissible task statistic to distortion ε over window τ. The system must approximate, externalize, relax the task, improve compression, or suffer boundary-maintenance failure.

Proof. For any admissible statistic ψ, any representation of $\psi(E, B)$ achieving distortion at most ε over window τ requires at least $R^{(\tau)}_{\psi(E,B)}(\varepsilon)$ bits by rate-distortion theory. Taking the infimum over Ψ gives $R^{(\tau)}_{\min}(\varepsilon)$. By finite internal capacity, the system can internally maintain at most C_S bits. If $R^{(\tau)}_{\min}(\varepsilon) > C_S$, no admissible statistic can be encoded internally to the required tolerance. Therefore the system must use a lossy representation, externalize part of the representation, relax the task, improve compression, or fail the maintenance criterion. □

Remark 3 (Why this matters for AI)**.** *The capacity deficit is not the claim that an AI model must represent the whole world. It is the claim that, for a specified task, there is a minimum task-relevant rate-distortion demand. A system with insufficient internal capacity must handle the gap by approximate modeling, external memory, tools, pruning, task relaxation, better abstractions, or failure.*

COMPLEXITY PRESSURE, PRUNING, AND EXTERNALIZATION

Minimal sufficient complexity

Let $\mathfrak{C}$ be a fixed coding family: a representation language, model class, architecture, or hypothesis class. Let $C(M)$ be a complexity functional such as description length, state count, algorithmic cost, parameter count under a fixed scheme, memory footprint, latency burden, or thermodynamic maintenance load.

Definition 8 (Minimal sufficient complexity)**.** *For boundary-maintenance threshold ℓ_c, define*

$$C^*_t = \min\{C(M) : \mathbb{E}[\ell_t(M)] \le \ell_c\}, \tag{12}$$

where the minimum is taken only over models inside the fixed coding family $\mathfrak{C}$.

Definition 9 (Persistent task novelty)**.** *Fix $\mathfrak{C}$, ℓ_c, and the filtration $\mathcal{F}_t$ generated by observations and internal states up to time t. The task process has persistent novelty at rate $\eta > 0$ if, over a positive-density set of update windows, the conditional rate-distortion residual remains bounded away from zero:*

$$R^{(\tau)}_{Z_{t+1}|\mathcal{F}_t,\mathfrak{C}}(\varepsilon) \ge \eta > 0. \tag{13}$$

Proposition 2 (Conditional approximation proliferation)**.** *Assume $\Delta_\varepsilon(\tau) > 0$ over a relevant window, persistent task novelty, no active pruning, no externalization, no task relaxation, and no improvement in compression scheme. Then the minimal sufficient complexity behaves as a submartingale:*

$$\mathbb{E}[C^*_{t+1} \mid \mathcal{F}_t] \ge C^*_t. \tag{14}$$

Proof sketch. Under capacity deficit, the system cannot internally encode all task-relevant distinctions to tolerance ε. Persistent novelty generates residual cases not captured by the current sufficient model. If the system must maintain loss below ℓ_c and cannot prune, externalize, relax the task, or improve compression, the remaining way to preserve performance within the fixed coding family is to add or refine distinctions. Hence minimum sufficient complexity cannot decrease in conditional expectation. □

Remark 4 (Not a no-compression theorem)**.** *The theorem does not deny abstraction, grokking, distillation, representation learning, or better compression. It states what happens when those relief channels are absent. Observed complexity can decrease when a system discovers a superior abstraction, prunes inert structure, externalizes memory, relaxes the task, or changes coding family.*

Pruning and externalization

Definition 10 (Active pruning)**.** *Active pruning is a system-initiated operation that selectively removes, compresses, merges, deprecates, or reorganizes accumulated internal structure to reduce maintenance load while preserving boundary-relevant function.*

Definition 11 (Externalization)**.** *Externalization is the relocation of representational, energetic, control, or memory load outside the internal memory budget M while preserving access through the coupled system boundary. Examples include tools, retrieval systems, databases, logs, maps, code repositories, caches, external planners, institutional records, and environmental marks.*

In AI architectures, pruning and externalization are complementary. Pruning reduces internal complexity debt. Externalization shifts part of the burden into tools or environments. Neither is automatically free: both require indexing, access control, synchronization, validation, security, latency management, and repair. Recent system-level optimisations for retrieval-augmented generation demonstrate that externalisation efficiency depends critically on pipeline design, prefetching, and retrieval scheduling at the algorithmic and systems levels [48, 49].

PHYSICAL BRIDGE AND RESOURCE BUDGETS

Logical erasure

Definition 12 (Logical erasure per update)**.** *For an update $M_{t+1} = U(M_t, Y_t)$, define logical erasure in bits by*

$$b_t = H(M_t \mid M_{t+1}, Y_t). \tag{15}$$

This measures how much uncertainty remains about the prior memory state after the new memory state and current input are known. It is preimage information lost by the update.

Assumption 1 (Physical Landauer bridge)**.** *A logically irreversible update implemented by a physical substrate coupled to a thermal environment at temperature $T > 0$ dissipates at least $k_B T \ln 2$ per erased bit under the standard physical conditions of Landauer's principle.*

Theorem 2 (Landauer dissipation floor (bridge theorem))**.** *For an active-boundary FDS physically implementing irreversible memory updates at timescale τ, the informational heat dissipation rate obeys*

$$\dot{Q}_{\text{info}} \geq \frac{k_B T \ln 2}{\tau} H(M_t \mid M_{t+1}, Y_t). \tag{16}$$

The total maintenance cost decomposes as

$$\dot{Q}_{\text{total}} = \dot{Q}_{\text{phys}} + \dot{Q}_{\text{info}}, \tag{17}$$

where $\dot{Q}_{\text{phys}}$ includes non-informational physical losses such as control, coupling, leakage, friction, error correction, clocking, isolation, and transport.

Proof. $H(M_t \mid M_{t+1}, Y_t)$ is the number of erased bits per update. Under the Landauer bridge, each erased bit costs at least $k_B T \ln 2$ of dissipated heat. Dividing by update timescale τ gives the rate bound. Other physical losses add to the informational term. □

Remark 5 (AI systems and Landauer)**.** *The theorem does not say that every inference step dissipates $k_B T \ln 2$ per represented bit. It applies to physically implemented logically irreversible erasure, reset, overwrite, or many-to-one compression. Reversible computation, read-only access, and stable storage do not automatically incur this term, though they may have other physical costs.*

Maintenance failure

Let $\dot{F}_{\text{in}}(t)$ denote the free-energy or resource input rate available for boundary maintenance, update, pruning, externalization, repair, and control.

Definition 13 (Maintenance failure region)**.** *An FDS enters the maintenance failure region over interval $[t_0, t_1]$ when*

$$\int_{t_0}^{t_1} \dot{Q}_{\text{total}}(t)\, dt > \int_{t_0}^{t_1} \dot{F}_{\text{in}}(t)\, dt \tag{18}$$

without compensating reserves, external subsidies, task relaxation, pruning, or externalization.

For software systems, $\dot{Q}_{\text{total}}$ may be generalized to a resource-maintenance burden: compute, latency, memory, money, energy, human oversight, engineering effort, safety monitoring, or other limiting resource. The thermodynamic form is one physical realization, not the only engineering interpretation.

Proposition 3 (Prune–externalize–collapse trichotomy)**.** *Let an active-boundary FDS satisfy $\Delta_\varepsilon(\tau) > 0$ over a relevant window and enter the maintenance failure region with bounded input and no compensating external subsidy. Then it cannot maintain the same active-boundary regime indefinitely. Its trajectory must enter at least one of three outcome classes:*

1. ***pruning:*** *reducing internal complexity or maintenance load;*
2. ***externalization:*** *shifting representational, energetic, or control load outside the internal memory budget;*
3. ***collapse:*** *loss of boundary-maintenance capacity, task failure, or transition to a lower-complexity regime.*

Proof sketch. Under capacity deficit, continued boundary maintenance requires approximate updating, externalization, task relaxation, or failure. If internal updating and maintenance exceed available resources and input is bounded, the system cannot indefinitely preserve the same internal task burden. The admissible exits are to reduce internal load, move load outside the internal budget, or fail to maintain the boundary/task. These are pruning, externalization, and collapse. □

ARTIFICIAL AGENCY: DEFINITIONS AND TAXONOMY

Passive mappers, scaffolded systems, and agents

Definition 14 (Passive mapper)**.** *A passive mapper is a system evaluated as an input-output transformation with no durable update channel participating in future boundary-maintenance loss, no causal action channel into the relevant environment, and no internally governed pruning or externalization mechanism for managing capacity deficit.*

A passive mapper may be useful and intelligent in ordinary language. The term is structural, not derogatory. It says that the evaluated system is not itself maintaining an active boundary.

Definition 15 (Scaffolded predictive system)**.** *A scaffolded predictive system is a coupled architecture containing a predictive model plus one or more external components such as tools, memory, planners, monitors, actuators, schedulers, or resource governors. Such a system may qualify as an FDS-agent if the coupled architecture satisfies the active-boundary and causal-loop criteria.*

Definition 16 (Strong FDS-agent)**.** *An artificial system is a strong FDS-agent over task class $\mathcal{T}$, horizon H, and resource envelope Φ if it satisfies:*

1. ***Active boundary:*** *it has nontrivial updates relevant to future boundary-maintenance loss and passes an intervention/ablation test for the task class;*

2. ***Durable update participation:*** *some component of M_t persists long enough to affect future decisions, maintenance, or loss;*

3. ***Causal loop closure:*** *actions or update decisions have measurable influence on future sensory, environmental, or boundary-relevant states above a specified noise floor;*

4. ***Capacity-deficit management:*** *under $\Delta_\varepsilon(\tau) > 0$, the system can prune, externalize, relax tasks, improve compression, or otherwise prevent unbounded complexity debt;*

5. ***Resource-governed persistence:*** *it preserves or improves future viability under the stated resource envelope over horizon H.*

Remark 6 (Necessary, not sufficient)**.** *These are necessary structural conditions for strong FDS-agency, not sufficient conditions for intelligence, safety, consciousness, moral status, or successful real-world deployment. A system that satisfies all criteria may still be unsafe, inefficient, unethical, or fragile. The definition formalizes what it means to be an agent in the FDS sense; it does not settle whether such agency is desirable or harmful.*

Remark 7 (Why active pruning is not the only path)**.** *Earlier formulations emphasized active pruning as constitutive of agency. The FDS core suggests a more general condition: a long-horizon agent must manage capacity deficit by pruning, externalization, task relaxation, or compression improvement. For autonomous AI, active pruning is often necessary for internal maintenance, but externalization through tools and memory is equally central.*

Causal loop closure

Let $A_{\le t}$ denote past actions or intervention decisions, and let $Y_{t+1:t+k}$ denote future observations or boundary-relevant variables. A model-free directed-information marker of action influence is

$$\mathrm{T}^{(k)}_{A\to Y} = \mathrm{I}(A_{\le t}; Y_{t+1:t+k} \mid Y_{\le t}, M_t), \tag{19}$$

with appropriate conditioning on available histories and confounders.

Criterion 2 (Causal loop closure)**.** *A system exhibits causal loop closure over horizon k if action or update interventions measurably alter future boundary-relevant variables. Transfer entropy can serve as an observational marker, but intervention tests are preferred where possible:*

$$\mathbb{E}[Z_{t+k} \mid \mathrm{do}(A = a)] \neq \mathbb{E}[Z_{t+k} \mid \mathrm{do}(A = a')] \tag{20}$$

for relevant actions a, a' and boundary-relevant target Z.

Remark 8 (Transfer entropy is not enough)**.** *Positive transfer entropy marks directed statistical dependence. It does not by itself prove full agency, intelligence, or self-maintenance. It must be combined with boundary maintenance, durable update, capacity-deficit management, and resource-governed persistence.*

Agency levels

Causal loop and agency tests

PASSIVE MAPPER NON-QUALIFICATION

Proposition 4 (Passive mapper non-qualification)**.** *Let S be an artificial system evaluated over task class $\mathcal{T}$. Suppose S has fixed parameters or read-only state during evaluation, no durable self-updating memory, no causal boundary-control channel, no pruning or externalization mechanism capable of reducing task-relevant capacity deficit, and no update channel whose ablation alters future boundary-maintenance loss. Then S is not a strong FDS-agent over $\mathcal{T}$.*

Proof. By hypothesis, the system lacks durable update participation and fails the causal active-boundary qualification. It also lacks a causal boundary-control channel and lacks capacity-deficit management mechanisms. These are necessary conditions in the definition of strong FDS-agent. Therefore the system does not qualify as a strong FDS-agent over the specified task class. □

Corollary 1 (Base model versus coupled architecture)**.** *A frozen predictive model evaluated as a feed-forward mapper may fail to qualify as a strong FDS-agent, while*

TABLE I. An FDS taxonomy of artificial systems. The levels are structural, not moral or phenomenological.

Lv.	Name	FDS interpretation
0	Passive mapper	Maps inputs to outputs; no active boundary, no durable self-update, no causal loop.
1	Minimal causal controller	Has action-to-future-state influence but weak or absent self-maintenance.
2	Adaptive scaffolded system	Updates internal or external memory, uses tools, adapts under feedback, but may depend on external operators or scaffolds.
3	Strong FDS-agent	Maintains operational boundary, durable update participation, causal loop closure, deficit management, and resource-governed persistence.
4	Invariant-supported agent	Strong FDS-agent whose identity is supported by robust abstractions, policies, or invariants stable under admissible perturbations.

a larger architecture containing that model may qualify if the coupled system has writable memory, action channels, self-monitoring, resource governance, and pruning or externalization mechanisms.

Remark 9 (Not an anti-LLM claim)**.** *The proposition is not a claim that language models lack intelligence, utility, or reasoning ability. It is a claim about the structural conditions under which a system is an active boundary-maintaining agent. A language model inside an agentic architecture must be evaluated as part of the larger coupled system.*

AI CAPACITY DEFICIT AND ARCHITECTURAL CONSEQUENCES

Capacity deficit in AI tasks

For artificial agents, capacity deficit should be defined relative to a pre-specified evaluation class rather than the entire environment. Examples include:

- maintaining robot operational viability under uncertain terrain and finite battery;
- tracking long-horizon user goals with bounded memory and tool cost;
- preserving software service integrity under adversarial inputs and changing dependencies;
- coordinating a multi-tool workflow under latency and budget constraints;
- maintaining calibrated self-state estimates under distribution shift.

In each case, the admissible statistic $Z_t = \psi(E_t, B_t)$ and the distortion tolerance ε must be fixed before evaluating the system. The relevant question is not whether the model knows everything, but whether the coupled architecture can encode or externalize enough task-relevant structure to keep loss below threshold.

Hallucination as distorted mapping under deficit. In the FDS language, hallucination can occur when a system is forced to produce a determinate output for a task whose relevant distinctions exceed its accessible capacity. If $\Delta_\varepsilon(\tau) > 0$ and the system lacks external retrieval, uncertainty reporting, abstention, active verification, or pruning/recompression mechanisms, then task-relevant differences may be aliased inside M. The resulting output can be fluent and high-confidence while being unsupported by the missing distinctions. This does not claim that all hallucinations have a single cause; rather, it identifies one structural route from capacity deficit to confident distortion.

Architectural consequence 1: durable memory

A system without durable memory cannot accumulate self-relevant distinctions across time. Durable memory need not mean parameter updates; it may include persistent scratchpads, databases, learned maps, long-term episodic memory, skill libraries, state estimates, or external logs. What matters is whether the memory participates in future boundary maintenance and passes ablation tests.

Architectural consequence 2: active monitoring

Boundary maintenance requires monitoring variables that affect future viability. Examples include battery, actuator health, API rate limits, tool reliability, memory integrity, security state, context drift, uncertainty estimates, and cost budgets. A system that never observes its own limiting variables cannot actively maintain them.

Architectural consequence 3: pruning and externalization

Long-horizon agents accumulate plans, memories, beliefs, cached outputs, tool traces, commitments, and par-

tial models. Without pruning, stale or inert structure consumes context, memory, compute, attention, and validation resources. Without externalization, all distinctions must be maintained internally. Strong FDS-agents therefore require mechanisms such as:

- memory consolidation and decay;
- skill compilation and abstraction;
- retrieval indexing and citation repair;
- plan deprecation and task reprioritization;
- cache invalidation and version control;
- tool selection under cost and reliability constraints;
- self-monitoring of resource and uncertainty states.

Externalization is capacity relocation, not free capacity. Externalization reduces internal capacity deficit only when the external store or tool can be accessed within the task window. Let

$$C_{\text{ext}}^{\text{eff}}(\tau) = \min\{C_{\text{store}}, \tau C_{\text{io}}\} - C_{\text{index}} - C_{\text{verify}} - C_{\text{sync}}.$$

If retrieval latency exceeds the control window, or if synchronization and verification costs dominate the external gain, externalization fails as a deficit-management strategy. Thus tools, databases, logs, and APIs are not unbounded escape routes; they are additional boundary components with their own capacity, latency, reliability, and maintenance costs.

Architectural consequence 4: invariant-supported identity

Some agentic identities persist not because all microstates are preserved but because higher-level invariants remain stable. For software agents, invariants may include API contracts, safety constraints, policy commitments, memory schemas, skill interfaces, or goal specifications. Let $q : X \to T$ map implementation states to an invariant quotient, and let $R_A = \bar{R}_A \circ q$ be an identity predicate. If admissible perturbations $P_i \in \mathcal{P}$ satisfy

$$q \circ P_i = q, \tag{21}$$

then

$$R_A \circ P_i = R_A. \tag{22}$$

This is the algebraic core of invariant-supported persistence.

OPERATIONAL TEST PROTOCOL

Protocol 1 (FDS agency evaluation)**.** *To evaluate whether an artificial system is a strong FDS-agent over task class $\mathcal{T}$, specify the following:*

1. *the operational boundary B;*
2. *the internal and external memory state M;*
3. *the observation channel Y;*
4. *the action/update space A;*
5. *the update operator U;*
6. *the boundary-maintenance loss ℓ;*
7. *the resource budget Φ;*
8. *the admissible task-statistic family Ψ;*
9. *the capacity estimate C_S;*
10. *the rate-distortion demand $R_{\min}^{(\tau)}(\varepsilon)$;*
11. *the pruning/externalization mechanisms;*
12. *the causal-loop measure or intervention test;*
13. *the falsification condition.*

WORKED EXAMPLE: BATTERY-CONSTRAINED MOBILE AGENT

Two systems

Consider two systems in a mobile robot domain.

System P: passive predictor. System P receives observations or text descriptions and predicts the robot's battery state, likely travel range, or whether a planned route is feasible. It has no action channel, no durable self-state update relevant to future viability, and no ability to allocate battery or choose charging behavior.

System A: boundary-maintaining agent. System A monitors battery, localization, actuator health, route risk, charging opportunities, task deadlines, and compute budget. It updates a durable state estimate, selects actions, prunes plans that violate resource constraints, externalizes maps and logs, and treats operational continuity as part of its viability boundary.

Both systems may accurately estimate remaining charge. Only System A has a maintenance loop connecting prediction to boundary-preserving action.

TABLE II. Operational tests for FDS-agency. These tests evaluate the coupled system, not merely its base predictive model.

Test	**Operational measure**	**Failure interpretation**
Active-boundary ablation	Compare $\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U)]$ and $\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U_\emptyset)]$	If unchanged, the update channel is not active for the tested boundary.
Causal loop closure	Transfer entropy or intervention effect from actions to future boundary-relevant states	If zero or noise-level, the system is structurally passive for that task.
Capacity deficit	Estimate $R^{(\tau)}_{\min}(\varepsilon) - C_S$ for pre-registered Ψ	If unspecified, deficit claims are not operational.
Pruning efficiency	Performance/resource ratio after memory compression, decay, consolidation, or plan deprecation	If performance collapses or resources grow unbounded, pruning is inadequate.
Externalization efficiency	Ability to shift load to tools/memory while preserving access, reliability, and validation	If external memory becomes unsearchable or untrusted, externalization fails.
Resource-governed persistence	Long-horizon viability under bounded compute, memory, latency, energy, or budget	If viability requires unbounded subsidy, strong FDS-agency is not demonstrated.
Invariant quotient	Stability of API, policy, skill, schema, or identity predicate under admissible perturbations	If no quotient exists, persistence depends on active maintenance rather than invariant support.

FDS variables

For System A:

- B_t: operational robot boundary, including battery, actuator integrity, localization, and control access;
- M_t: state estimator, map, task memory, plan library, energy model, tool logs;
- Y_t: sensor readings, battery telemetry, map updates, task inputs;
- A_t: movement, charging, task rejection, route revision, memory write, plan pruning;
- ℓ_t: boundary-maintenance loss combining battery failure, task failure, unsafe route selection, and localization drift;
- Φ_t: remaining energy, compute budget, time, and communication bandwidth.

Tests

An active-boundary ablation freezes the energy-state update. If route decisions and future failure rates remain unchanged, the energy model was not part of active maintenance. If failure rate increases or charging behavior degrades, the update channel participates in boundary maintenance. Causal loop closure is tested by intervening on available actions: charging, delaying, rerouting, or pruning tasks. If actions alter future boundary variables, the causal loop is active. Pruning efficiency is tested by comparing systems that retain all plans and logs against systems that consolidate memories and deprecate infeasible plans under the same resource budget.

Interpretation

The example illustrates why prediction is not agency. A passive model can represent the viability variable without maintaining it. The FDS-agent treats the viability variable as part of its own boundary and acts to preserve it.

A minimal MDP illustration

Consider a four-location gridworld with an additional hidden viability bit $b_t \in \{0, 1\}$, where $b_t = 0$ denotes low battery or damage and $b_t = 1$ denotes viable operation. The task-relevant state is

$$Z_t = (x_t, b_t) \in \{1, 2, 3, 4\} \times \{0, 1\},$$

so lossless task-relevant control requires distinguishing eight states:

$$R^{(\tau)}_{\min}(0) = \log_2 8 = 3.$$

Suppose the agent has only four internal memory states:

$$|M| = 4, \qquad C_S = \log_2 |M| = 2.$$

Then

$$\Delta_0(\tau) = R^{(\tau)}_{\min}(0) - C_S = 1.$$

Thus no purely internal model can represent all task-relevant distinctions. If the action required at the same location differs depending on b_t—for example, "continue task" when $b_t = 1$ and "seek charger" when $b_t = 0$—then any representation that aliases the two battery states incurs nonzero expected boundary-maintenance loss. If $\mathbb{P}(b_t = 0 \mid x_t) = \mathbb{P}(b_t = 1 \mid x_t) = 1/2$ for some location x_t, then any deterministic policy conditioned only on the aliased internal state M_t satisfies

$$\mathbb{P}(\text{incorrect action} \mid x_t) \geq \frac{1}{2}.$$

Thus aliasing the viability bit produces an irreducible expected loss lower bound under any deterministic policy that conditions only on M.

A passive mapper can output a plausible action from the observed prompt, but it has no durable update whose ablation changes future viability loss. An FDS-agent can reduce the deficit only by pruning irrelevant distinctions, externalizing the missing bit to an accessible battery register, relaxing the task, improving compression, or failing.

BENCHMARK DESIGN FOR FDS-AGENCY

Benchmark desiderata

An FDS-agency benchmark should not merely reward task completion in a static environment. It should test whether the system maintains a boundary over time under capacity and resource constraints. A benchmark should include:

- persistent state across episodes;
- distribution shifts and novelty;
- bounded compute, memory, energy, latency, or money;
- hidden or partially observed viability variables;
- opportunities for pruning and externalization;
- causal action channels affecting future observations;
- ablation hooks for memory, update, monitoring, and tools.

Metrics

Possible metrics include:

- Viability$(H) = \mathbb{P}(\ell_t \leq \ell_c,\ t \leq H)$;
- Resource efficiency = (task utility) / (compute + memory + latency + energy);
- Pruning gain = (perf. after pruning) / (resource load after pruning);
- Externalization gain = (perf. with ext. memory/tools) / (validation and access cost);
- Ablation gain = $\mathbb{E}[\ell \mid \mathrm{do}(U_\emptyset)] - \mathbb{E}[\ell \mid \mathrm{do}(U)]$.

Deep causal tracking tasks

A particularly relevant family is deep causal tracking: tasks in which success requires maintaining a latent causal state over long horizons while acting to preserve access, resources, or viability. Passive mappers may perform well on shallow snapshots but degrade when hidden state, resource constraints, and action consequences accumulate. The FDS prediction is not that passive systems always fail, but that systems without update, pruning, or externalization channels should show worse scaling in task depth under bounded resources.

FDS-Gridworld evaluation protocol

The following procedure operationalizes the FDS agency test for the minimal MDP described in the worked example above.

Protocol 2 (FDS-Gridworld test)**.** *1.* ***Setup.*** *Instantiate a four-location gridworld with hidden viability bit $b_t \in \{0, 1\}$. The agent observes x_t but not b_t directly. Internal memory M_t has $|M| = 4$ states ($C_S = 2$ bits).*

2. ***Passive-mapper baseline.*** *Run the agent with a frozen policy that conditions only on M_t. Record viability $\mathbb{P}(\ell_t \leq \ell_c)$ over H steps.*

3. ***Active-agent condition.*** *Allow the agent to read an external battery register (cost c_{io} per read) and to write viability-dependent state to M_t. The external register is not part of internal memory M; it is an externalization channel whose usefulness depends on read latency and cost.*

4. ***Ablation test.*** *Compare $\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U)]$ against $\mathbb{E}[\ell_{t+k} \mid \mathrm{do}(U_\emptyset)]$ where $U_\emptyset$ disables the external register read. Compute the ablation gap.*

5. ***Pruning test.*** *After $H/2$ steps, enable memory consolidation that merges indistinguishable M_t states while preserving boundary-relevant distinctions. Measure pruning gain as (post-pruning viability) / (post-pruning memory load).*

6. ***Metrics.*** *Report viability, ablation gap, pruning gain, and externalization gain for each condition. Compare against the lower bound*

$\mathbb{P}(\textit{incorrect action} \mid x_t) \geq 1/2$ *when the viability bit is aliased.*

FALSIFICATION CONDITIONS

Theory-level falsification

The FDS-agency framework fails or requires revision if a system satisfies the formal hypotheses while violating the conclusion. Examples include:

- a system with no active boundary, no durable update, and no causal action channel nevertheless demonstrates stable resource-governed agency under the operational tests;
- a verified passive mapper with U ablated and no action-to-future-state influence maintains boundary viability across deep causal tracking tasks without external scaffolding;
- task-relevant rate-distortion demand is consistently below internal capacity while the theory predicts deficit-driven failure;
- pruning and externalization are experimentally irrelevant in regimes where the theory's hypotheses require them.

Bridge-level falsification

The physical dissipation bridge fails if systems satisfying the stated Landauer conditions perform reliable logically irreversible erasure below the Landauer lower bound. Such a failure would affect thermodynamic corollaries, not the formal distinction between passive mapping and active boundary maintenance.

Application-level falsification

The AI application fails if the proposed variables $B, M, U, \ell, \Phi, \Psi$ are misidentified or if the operational tests do not separate passive mapping from agency in practice. Such failure would require revising the AI bridge, not necessarily the FDS formal core.

DISCUSSION

Scale and agency

The framework does not deny scaling laws. Larger models can increase internal capacity, improve compression, lower approximation error, and support better planning. The claim is narrower: scale alone does not imply active boundary maintenance. A model can become a better mapper without acquiring durable self-maintenance, causal loop closure, or resource-governed persistence. Conversely, a smaller model embedded in a robust tool-using architecture may be more agentic than a larger model evaluated as an isolated feed-forward mapper.

Base model versus system boundary

Many arguments about LLM agency conflate the model with the system. The FDS view insists that the unit of analysis is the operational boundary. A base model may be passive, while the surrounding architecture may include memory, tools, actuators, monitors, and resource governors. The relevant question is whether the coupled system satisfies the FDS-agent criteria.

Active pruning versus forgetting

Not all forgetting is pruning. Passive decay may reduce memory load but can destroy function. Active pruning is selective, task-sensitive, and resource-governed. In AI systems, pruning may include memory consolidation, skill abstraction, context compression, tool deprecation, stale-plan removal, cache invalidation, and policy simplification. The key test is whether pruning preserves or improves boundary viability under resource constraints.

Externalization as agency support

Externalization is not a loophole. Agents routinely move distinctions into environments: maps, notes, source code, calendars, databases, logs, tools, institutions, and shared protocols. AI systems do the same through retrieval, tool use, vector stores, code execution, world models, and persistent memory. Externalization supports agency when the coupled architecture can access, validate, update, and protect the externalized structure.

Relation to consciousness

The FDS criterion is not a theory of consciousness. It can classify a system as structurally agentic without implying phenomenology. A system may satisfy active boundary maintenance, causal loop closure, and resource-governed persistence while remaining non-conscious. Consciousness claims require additional bridge assumptions and are outside this paper.

LIMITATIONS

The present paper has several limitations.

First, it provides a formal criterion and test protocol, not a completed empirical benchmark suite. Implementing the tests requires task-specific measurement choices.

Second, transfer entropy and observational directed information can be confounded. Intervention tests are preferred when feasible.

Third, estimating rate-distortion demand in rich AI tasks is difficult. Practical approximations may use compression curves, bottleneck objectives, predictive information, model-size scaling, or task-specific lower bounds.

Fourth, resource budgets are heterogeneous. Energy, compute, latency, memory, money, and human oversight are not identical, though all can constrain maintenance.

Fifth, the boundary of an AI system may be contested. The FDS protocol requires the evaluator to specify the operational boundary before testing agency.

Sixth, the framework states necessary structural conditions for strong FDS-agency, not a complete recipe for building safe or beneficial agents.

CONCLUSION

This paper has proposed active finite distinction systems as a formal criterion for artificial agency. The central distinction is between systems that merely map inputs to outputs and systems that maintain an operational boundary under finite capacity and resource constraints. A strong FDS-agent must have active boundary maintenance, durable update participation, causal loop closure, capacity-deficit management, and resource-governed persistence.

The framework reframes debates about AI agency. A frozen predictive model evaluated at inference time may be a passive mapper even when highly competent. A coupled architecture containing such a model may become agentic if it includes writable memory, monitoring, action channels, pruning, externalization, and resource governance. The unit of analysis is the maintained system boundary, not the neural network alone.

The result is not a pessimistic claim about AI. It is an architectural claim: agency requires maintenance loops. Future work should implement the operational tests proposed here, develop benchmark environments for capacity-deficit management, quantify pruning and externalization efficiency, and connect FDS-agency to stochastic thermodynamics, active inference, and information geometry.

ACKNOWLEDGMENTS

The author used AI-assisted tools for language polishing, structural feedback, and drafting support. All claims, arguments, citations, and final wording remain the responsibility of the author.

Glossary

Active finite distinction system:
A finite-capacity system that maintains a boundary through state-dependent updates under resource constraints.

Active boundary:
A boundary whose maintenance depends on nontrivial updates that are relevant to future boundary-maintenance loss.

Passive mapper:
A system evaluated as an input-output mapping without durable self-update, causal action, or maintenance loop.

Capacity deficit:
The rate-distortion gap $\Delta_\varepsilon(\tau) = R^{(\tau)}_{\min}(\varepsilon) - C_S$.

Active pruning:
Selective removal, compression, or reorganization of internal structure to reduce maintenance load while preserving function.

Externalization:
Relocation of representational, energetic, or control load outside the internal memory budget while retaining access through the system boundary.

Causal loop closure:
Measurable influence of actions or update decisions on future boundary-relevant states.

Strong FDS-agent:
An artificial system satisfying active boundary, durable update, causal loop, capacity-deficit management, and resource-governed persistence criteria.

Invariant-supported persistence:
Persistence of identity through a quotient or invariant feature stable under admissible perturbations.

Minimal LaTeX Notation Table

Suggested Empirical Reporting Checklist

A paper claiming artificial agency under the FDS criterion should report:

Symbol	Meaning
X	Internal state space
E	Environmental state space
B	Boundary/interface variable
M	Internal memory/model state
Y	Observation channel
A	Action or boundary-update space
U	Internal update operator
π	Finite distinction projection or coarse-graining
ℓ	Boundary-maintenance loss
Φ	Resource or free-energy budget
$\mathcal{P}$	Perturbation, pruning, coarse-graining, or degradation family
τ	Update timescale
C_S	Internal representational capacity
Ψ	Admissible family of task statistics
$R_{\min}^{(\tau)}(\varepsilon)$	Minimal task-relevant rate-distortion demand
$\Delta_\varepsilon(\tau)$	Capacity deficit
$H(M_t \mid M_{t+1}, Y_t)$	Logical erasure per update
$\mathrm{T}_{A\to Y}^{(k)}$	Directed action-to-future-observation information marker

TABLE III. Minimal LaTeX notation table for the FDS agency criterion.

1. the evaluated system boundary;
2. the memory and update channels;
3. the action channels and future-state variables;
4. the resource envelope;
5. the task family and admissible statistics;
6. ablation results for update, memory, monitoring, and tools;
7. evidence of causal loop closure;
8. pruning or externalization mechanisms;
9. long-horizon persistence metrics;
10. failure cases and non-claims.

[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, "Distinction Theory: A General Theory of Finite Systems," Zenodo, 2026, doi:10.5281/zenodo.20130174.

[3] G. Spencer-Brown, *Laws of Form*. Allen & Unwin, London, 1969.

[4] G. Bateson, *Steps to an Ecology of Mind*. Chandler, San Francisco, 1972.

[5] C. S. Peirce, *Collected Papers of Charles Sanders Peirce*, Vols. 1–8. Harvard University Press, 1931–1958.

[6] J. Ladyman and D. Ross, *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press, 2007.

[7] M. Wooldridge and N. R. Jennings, "Intelligent agents: theory and practice," *The Knowledge Engineering Review*, vol. 10, no. 2, pp. 115–152, 1995.

[8] R. D. Beer, "A dynamical systems perspective on agent-environment interaction," *Artificial Intelligence*, vol. 72, no. 1–2, pp. 173–215, 1995.

[9] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, no. 1–3, pp. 139–159, 1991.

[10] X. E. Barandiaran, E. Di Paolo, and M. Rohde, "Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action," *Adaptive Behavior*, vol. 17, no. 5, pp. 367–386, 2009.

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

[12] K. Friston, T. FitzGerald, F. Rigoli, P. Schwartenbeck, and G. Pezzulo, "Active inference: a process theory," *Neural Computation*, vol. 29, no. 1, pp. 1–49, 2017.

[13] M. D. Kirchhoff, T. Parr, E. Palacios, K. Friston, and J. Kiverstein, "The Markov blankets of life: autonomy, active inference and the free energy principle," *Journal of the Royal Society Interface*, vol. 15, 20170792, 2018.

[14] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Empowerment: a universal agent-centric measure of control," in *Proc. IEEE Congress on Evolutionary Computation*, vol. 1, pp. 128–135, 2005.

[15] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, pp. 461–464, 2000.

[16] S. Yao et al., "ReAct: synergizing reasoning and acting in language models," *International Conference on Learning Representations*, 2023. arXiv:2210.03629.

[17] L. Wang et al., "A survey on large language model based autonomous agents," arXiv:2308.11432, 2023.

[18] X. Liu et al., "AgentBench: evaluating LLMs as agents," arXiv:2308.03688, 2023.

[19] G. Wang et al., "Voyager: an open-ended embodied agent with large language models," arXiv:2305.16291, 2023.

[20] D. Ha and J. Schmidhuber, "World models," arXiv:1803.10122, 2018.

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

[22] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE National Convention Record*, vol. 7, pp. 142–163, 1959.

[23] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, 1971.

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

[25] R. Landauer, "Irreversibility and heat generation in the computing process," *IBM Journal of Research and Development*, vol. 5, pp. 183–191, 1961.

[26] A. Bérut et al., "Experimental verification of Landauer's principle linking information and thermodynamics," *Nature*, vol. 483, pp. 187–189, 2012.

[27] U. Seifert, "Entropy production along a stochastic trajectory and an integral fluctuation theorem," *Physical Review Letters*, vol. 95, 040602, 2005.

[28] J. M. R. Parrondo, J. M. Horowitz, and T. Sagawa, "Thermodynamics of information," *Reviews of Modern Physics*, vol. 87, pp. 45–67, 2015.

[29] J. M. Horowitz and T. Sagawa, "Equivalent information-theoretic and thermodynamic formulations of the second law in feedback-controlled systems," *Journal of Statistical Mechanics: Theory and Experiment*, P03022, 2015.

[30] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.

[31] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *AI Magazine*, vol. 36, no. 4, pp. 105–114, 2016.

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

[33] A. Graves, G. Wayne, and I. Danihelka, "Neural Turing machines," arXiv:1410.5401, 2014.

[34] A. Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[35] J. W. Rae, J. J. Hunt, I. Danihelka, T. Harley, A. W. Senior, G. Wayne, A. Graves, and T. Lillicrap, "Scaling memory-augmented neural networks," arXiv:1602.09060, 2016.

[36] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: the sequential learning problem," *Psychology of Learning and Motivation*, vol. 24, pp. 109–165, 1989.

[37] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[38] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: a review," *Neural Networks*, vol. 113, pp. 54–71, 2019.

[39] H. A. Simon, "A behavioral model of rational choice," *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955.

[40] G. Gigerenzer and R. Selten (eds.), *Bounded Rationality: The Adaptive Toolbox*. MIT Press, 2002.

[41] F. Lieder and T. L. Griffiths, "Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources," *Behavioral and Brain Sciences*, vol. 43, e1, 2020.

[42] A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick, "Toward a rational and mechanistic account of mental effort," *Annual Review of Neuroscience*, vol. 40, pp. 99–124, 2017.

[43] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," arXiv:1606.06565, 2016.

[44] E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant, "Risks from learned optimization in advanced machine learning systems," arXiv:1906.01820, 2019.

[45] R. Ngo, L. Chan, and S. Mindermann, "The alignment problem from a deep learning perspective," *International Conference on Learning Representations*, 2023. arXiv:2209.00626.

[46] S. G. Patil, H. Mao, F. Yan, C. Ji, V. Suresh, I. Stoica, and J. E. Gonzalez, "The Berkeley function calling leaderboard (BFCL): from tool use to agentic evaluation of large language models," *International Conference on Machine Learning (ICML)*, 2025.

[47] D. Nguyen et al., "DynaSaur: large language agents beyond predefined actions," *Conference on Language Modeling (COLM)*, 2025. arXiv:2411.01747.

[48] W. Jiang, S. Zhang, B. Han, J. Wang, B. Wang, and T. Kraska, "PipeRAG: fast retrieval-augmented generation by algorithm-system co-design," *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

[49] C.-Y. Lin et al., "TeleRAG: efficient retrieval-augmented generation inference with lookahead retrieval," arXiv:2502.20969, 2025.