A Mathematical Theory of Agency and Intelligence

Wael Hafez
Semarx Research LLC
Alexandria, VA, USA
w.hafez@semarx.com

Abstract. Current AI systems achieve impressive task performance yet remain fragile, lacking the intrinsic ability to regulate the quality of their coupling with the environment. We introduce an information-theoretic framework that defines agency not by reward, but by Bi-Predictability (P)—a dimensionless measure of bidirectional coupling efficiency. From first principles, we establish that while passive physical systems are bounded at $P = 0.5$, agency inherently trades maximal coherence for the freedom to act. We define 'Intelligence' as the capacity to actively manage this trade-off via Self-Monitoring and Adaptation. Unlike prevailing paradigms that rely on sparse reward signals or internal prediction error—which often fail to register stability loss until task performance collapses—Bi-Predictability isolates coupling deviations in real-time, independently of task semantics. Validating this framework across chaotic physical systems, reinforcement learning, and large language models, we show that stability failures manifest as statistical deviations in P—signaling either decoherence (confusion) or pathological rigidity (fixation)—that remain invisible to standard performance metrics. To address this, we propose the Coupled Agency Architecture, which pairs a learning policy with an Information Digital Twin (IDT) capable of modulating interaction bandwidth via "reflexive wrappers." This mechanism, is inspired by biological thalamic modulation, provides a mathematically grounded information architectural requirement for robust, self-regulating agency, and motivates an IDT-guided regulation loop.

1 Introduction

Current AI systems can achieve impressive performance yet remain fragile as autonomous agents—dependent on external monitoring, retraining, and human validation. A central reason is a missing capability: today's systems have no principled way to assess the quality of their own interaction with the environment. They optimize task-specific objectives without measuring whether their coupling to the world is becoming reliable, brittle, or degraded.

We introduce an information-theoretic framework that addresses this gap. Its core construct is Bi-Predictability (P): a dimensionless measure of how tightly an agent and its environment mutually constrain one another through interaction. P does not quantify how much information is present or transmitted; it quantifies how effectively the interaction supports bidirectional predictability—how well each side can be specified only in relation to the other.

Bi-Predictability also exhibits distinct regime limits. Under our definition, classical physical interactions have an upper bound of 1/2 that reflects finite predictive capacity at the chosen description level; introducing action adds freedom and directionality, and this typically lowers P further (Fig. 1). We treat the quantum case as a conceptual anchor: an analogous construction can reach unity in maximally nonseparable quantum correlations, highlighting the contrast between nonseparability, measurement, and agentic flexibility. The key point for this paper is the agentic regime: when actions enter the loop, coherence is traded for flexibility.

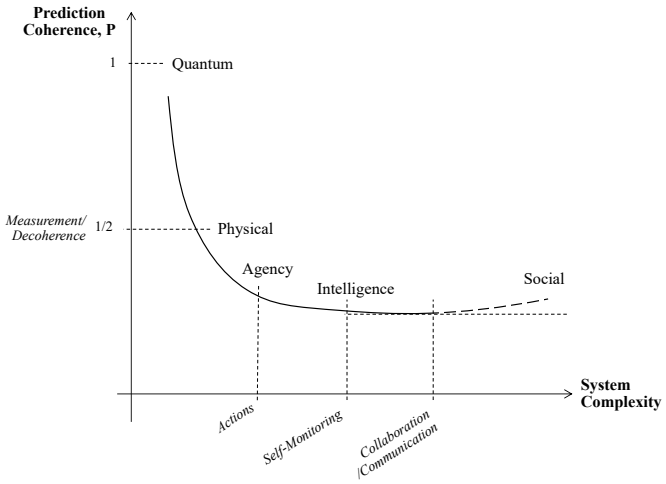

Figure 1 *Bi-Predictability cascade across interaction regimes. Conceptual trajectory of Bi-Predictability (P) as systems transition from physical interaction to agentic interaction. In classical physical systems, P is bounded under the present definition; when actions are introduced, added freedom and directionality typically reduce P at the level of an individual agent. Higher-level organization can partially recover or stabilize P by constraining future states through shared structure (for example, coordination, norms, and engineered environments). The quantum label serves as a conceptual anchor for non-separability rather than as a mechanistic claim about AI. Axes are schematic*

From P we derive operational definitions. Agency requires choice in the form of actions, and predictive asymmetry. Intelligence requires three capacities: learning, self-monitoring, and adaptation. Current AI largely provides learning and can exhibit agency, but it does not explicitly self-monitor interaction quality or adapt its own interface to preserve coupling under drift. As a result, scaling can increase throughput without indicating whether interaction quality is improving.

To address this gap, we propose the Coupled Agency Architecture, which pairs the learning policy with an Information Digital Twin (IDT). Unlike passive monitors, the IDT serves as a homeostatic regulator, computing Bi-Predictability from the observable interaction stream and modulating information efficiency—i.e., the effective shared predictability available at the observation–action interface—via adaptive interface-conditioning mechanisms. This mechanism, analogous to biological thalamic regulation of cortical signals, provides information architecture requirements and mathematically grounded blueprint for engineering robust, intelligent agency.

Shannon's theory formalized communication independent of meaning. Here we seek an analogous foundation for intelligent agency: measurable interaction-level quantities that define what an agent is, what it can do, and what current AI systems still lack.

2 Bi-Predictability (P)

We introduce a formal information-theoretic framework for quantifying how tightly two interacting entities constrain one another through their joint dynamics, independent of the absolute amount of information present or exchanged. Rather than asking how much information flows, the framework asks how much of the available uncertainty is shared—how mutually predictive the interaction is at the chosen level of description. Let S and S', denote successive states of the coupled system–environment interaction. We define Bi-Predictability, P as:

$$P = \frac{MI(S;S')}{H(S) + H(S')}$$

P measures the ratio of shared information to total information—not volume, but efficiency. $P = 1/2$ corresponds to ideal closed-loop interaction where states fully determine one another; $P = 0$ corresponds to successive states being statistically independent of one another.

2.1 The Universal Bound

From first principles, Bi-Predictability admits regime bounds. Under classical (Shannon) information, we obtain:

$$0 \leq P \leq \frac{1}{2}.$$

This bound is structural:

$$MI(S; S') \leq \min\left(H(S), H(S')\right)$$

so shared information cannot exceed half of the total entropy capacity $H(S) + H(S')$ under our definition. In the quantum setting, maximally nonseparable correlations can saturate the analogous construction (as highlighted by Bell-type phenomena, (Bell, 1964)), but the transition to classical definiteness—via measurement and decoherence—removes the correlations that permit P to approach unity.

2.2 Extension to Active Systems and Agency

The framework above applies to interacting entities in general. Many systems of interest, however—biological organisms and artificial agents—are active**:** they do not merely respond, but intervene. This introduces a natural asymmetry: one side maintains internal state and selects actions that influence what happens next.

We capture this by introducing an action variable A. Interaction is represented as $(S, A) \rightarrow S'$, where $S$is the agent's internal state (the information it uses to act), $A$is its chosen intervention, and $S'$is the resulting next state after the environment responds. Bi-Predictability generalizes to:

$$P = \frac{MI(S, A; S')}{H(S) + H(A) + H(S')}.$$

Under classical (Shannon) information, the same ceiling of $1/2$ applies in principle; in practice, introducing A makes this ceiling unattainable. Intuitively, action adds internal degrees of freedom that must be maintained while remaining coupled to the environment. Agents therefore trade maximal coherence for the ability to act.

Introducing Aalso makes predictability directional (Fig. 2). We define:

- **Forward predictive uncertainty** $H_f = H(S' \mid S, A)$: how uncertain outcomes remain given what the agent knew and did. High $H_f$indicates weak constraint of the environment's response by the agent's state and action.
- **Backward predictive uncertainty** $H_b = H(S, A \mid S')$: how many internal states and actions are consistent with an observed outcome. High H_b indicates many distinct causes collapsing to indistinguishable consequences.

Their difference defines a predictability asymmetry**:**

$$\Delta H = H_f - H_b,$$

which localizes how predictability is lost—whether primarily through environmental response uncertainty (high H_f) or through agent-side indistinguishability (high H_b). This decomposition matters: two systems can exhibit similar P yet fail for different reasons. P measures overall coupling efficiency; ΔH reveals where the coupling breaks.

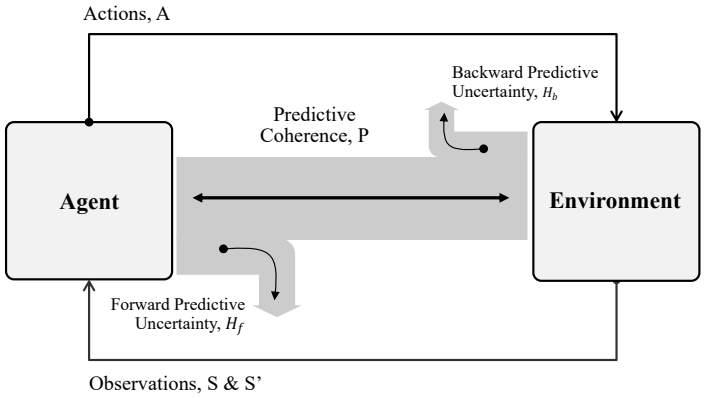

Figure 2 *Bi-Predictability and predictability imbalance in agent–environment interaction. An agent and environment form a closed interaction loop. Psummarizes coupling strength; ΔHindicates whether predictability breaks down mainly in the forward direction (environmental response) or backward direction (agent-side indistinguishishability.*

2.3 Interpretation

P and ΔH expose the trade-off introduced by agency: acting adds freedom, but intelligent action requires outcomes to be both controllable (low forward uncertainty) and legible (low backward ambiguity). Similar P values can hide different failure modes; ΔH separates them by indicating whether predictability is lost mainly in the environment's response (H_f) or in the agent's own indistinguishability (H_b).

P is not a normalization in the statistical sense—the denominator exceeds the numerator's theoretical maximum. P measures informational yield relative to total deployed capacity, not proximity to perfect coupling.

3 Physical Baseline (Double Pendulum)

We first test the framework on a deterministic physical system without an action channel: the double pendulum. This establishes a calibration point in which any loss of predictability arises from measurement/representation rather than decision-making.

3.1 Results

We analyzed two batches of 300 simulations spanning symmetric (equal mass) and asymmetric (unequal mass) settings.

Prediction 1: High Bi-Predictability. Under deterministic dynamics with a complete state representation, Pshould approach the classical ceiling of $1/2$. Consistent with this prediction, P remains close to the bound across both batches with low variance (Table 1), indicating that successive states are strongly mutually predictive despite chaotic sensitivity.

Batch	P Min	P Mean	P Max	P STD
1st batch	0.472944657	0.475747126	0.48095266	0.00155264
2nd batch	0.475749647	0.472919123	0.4769658	0.00169209

Table 1 *Bi-Predictability in a deterministic physical system. Summary statistics of P across double-pendulum simulations, showing values approaching the classical bound of 1/2.*

Prediction 2: Predictive asymmetry ≈ 0. In the absence of intervention or intrinsic randomness, forward and backward predictive uncertainty should be comparable, yielding $\Delta H \approx 0$. As predicted, forward and backward uncertainties are numerically indistinguishable and ΔH is centered near zero across both batches (Table 2).

Batch	Metric	Min	Mean	Max	STD
1st batch	Forward Predictive Uncertainty, $H(S'\|S)$	0.101762767	0.173011607	0.21491649	0.02207975
	Backward Predictive Uncertainty $H(S\|S')$	0.101853985	0.172982039	0.21480427	0.02213793
	Predictive Asymmetry $\Delta H = H(S'\|S) - H(S\|S')$	-3.76E-04	-6.6996E-07	5.15E-04	0.00016409
2nd batch	Forward Predictive Uncertainty $H(S'\|S)$	7.62E-02	0.132628187	0.19915062	0.01919794
	Backward Predictive Uncertainty $H(S\|S')$	7.63E-02	0.132629277	0.19949738	0.01920498
	Predictive Asymmetry $\Delta H = H(S'\|S) - H(S\|S')$	-6.25E-04	-1.09E-06	0.00063149	0.00022369

Table 2 *Forward and backward predictive uncertainty. Summary statistics of H_f H_b, and their difference (ΔH), demonstrating predictive symmetry in the absence of agency.*

Prediction 3: Chaos does not imply asymmetry. Across high-chaos regimes, Premains stable and ΔHremains near zero, supporting the distinction between chaotic sensitivity and directional loss of predictability in this setting.

3.2 Interpretation

Together, these results establish a physical calibration: for a deterministic system without an action channel, P approaches the classical ceiling and ΔH remains near zero. The small gap from the theoretical maximum is consistent with finite estimation and representation effects (for example, discretization and windowing) rather than dynamical limitations. In later agentic settings, departures from this pattern indicate that predictability is being lost through intervention and/or openness at the chosen interface.

4 Information Architecture of Agency and Intelligence

4.1 Agency (definition)

A system exhibits **agency** when an action variable Asatisfies three conditions:

- **Choice:** $H(A \mid S) > 0$— actions are not fully determined by the available state.
- **Effect:** $MI(A; S' \mid S) > 0$— actions change what happens next beyond what the state already predicts.
- **Predictive asymmetry:** $\mid \Delta H \mid > 0$— forward and backward predictive uncertainty differ.

Choice and effect are structural: the system can select among alternatives, and those alternatives matter. Predictive asymmetry is diagnostic: it indicates directional intervention at the (S, A, S') interface.

In deterministic physical dynamics without an action channel, forward and backward uncertainty remain balanced ($\Delta H \approx 0$) at the chosen description level, even under chaos. The double pendulum provides this baseline. Introducing action typically breaks this balance: outcomes do not fully "round-trip" back to the agent's internal causes, producing a measurable asymmetry in the predictive structure.

4.2 Intelligence (definition)

Agency enables intervention; **intelligence** manages the quality of that intervention. We define intelligence as requiring three capacities:

- **Learning:** increase overall interaction predictability $MI(S, A; S')$ (the numerator of P).
- **Self-monitoring:** measure and regulate Pover time.
- **Adaptation:** expand or reorganize the state, action, and outcome spaces $\{S\}, \{A\}, \{S'\}$—that is, change what the system can represent, what it can do, and what outcomes it can reliably bring about.

Learning builds coupling within a fixed interface; self-monitoring evaluates coupling efficiency; adaptation reshapes the interface itself.

By this definition, current AI typically achieves agency and learning, but lacks explicit self-monitoring and adaptation. Training can increase $MI(S, A; S')$ while leaving coupling efficiency unmeasured and $\{S\}, \{A\}, \{S'\}$ fixed by designers, which is why degradation detection still relies on external evaluation rather than first-person monitoring.

4.3 The Information Digital Twin (IDT)

To enable self-monitoring, we propose the Coupled Agency Architecture, which pairs the agentic policy with a regulatory Information Digital Twin (IDT) (Fig. 3). Unlike standard twins that replicate physical states, the IDT models interaction statistics, functioning as a homeostatic sidecar independent of the agent's internal model. The architecture operates in three stages: (1) Metric Estimation, where the IDT computes real-time P and ΔH from the (S, A, S') stream; (2) Stability Control, where a 'P Controller' detects statistical deviations from the coherent baseline; and (3) Reflexive Modulation, where significant excursions trigger interaction information efficiency modulation. By employing signal management techniques—such as action dampening ("Hold"), input filtering, or dimensionality reduction—this mechanism resolves open-loop fragility without requiring immediate retraining. In this way, the IDT supports real-time stability and provides further actionable insights for the adaptation of $\{S\}, \{A\}, \{S'\}$, to further improve agent Bi-Predictability and ultimately its decision effectiveness.

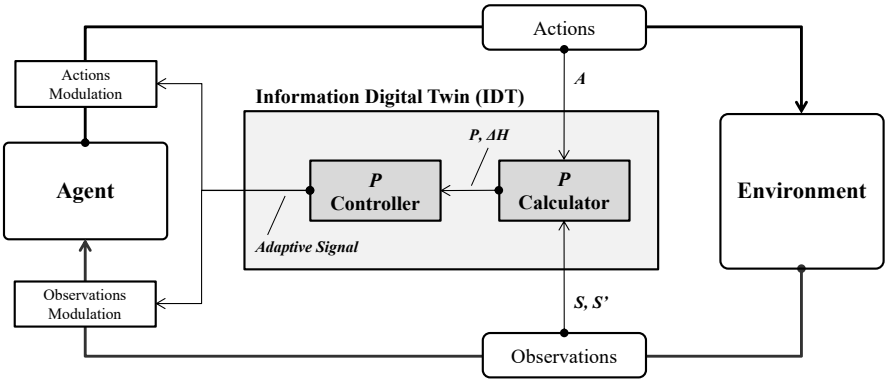

Figure 3 *Information Digital Twin architecture. All components—agent, environment, observations (S), actions (A), outcomes (S′)—are subject to noise and uncertainty. P captures the aggregate effect of these uncertainties on bidirectional coupling. The IDT receives copies of the (S, A, S') stream and computes P and ΔH. Modulation pathways (dashed) illustrate how these signals could regulate observation/action interfaces; closing this loop remains future work.*

By modulating the interface rather than the model weights, the system preserves agency during perturbations that would otherwise cause catastrophic drift. This functionally mirrors the mammalian thalamocortical loop, where thalamic nuclei monitor copies of sensory and motor signals and regulate signal transmission based on signal statistics rather than semantic content. The IDT thus provides the necessary engineering blueprint for converting passive predictive metrics into active, homeostatic agency.

4.4 Differentiation from Existing Frameworks

Existing frameworks define agency through feedback and stability (Wiener, 1948; Ashby, 1956), reward optimization robustness failures (Amodei et al., 2016; D'Amour et al., 2022; Shumailov et al., 2024), or intrinsic motivation such as empowerment (Schmidhuber, 1991; Klyubin et al., 2005) and prediction error minimization (Rao & Ballard, 1999; Friston, 2010). These approaches share a limitation: they measure unidirectional influence—agent → environment (empowerment) or environment → agent (prediction error)—not bidirectional coupling.

Bi-Predictability differs: P measures mutual coupling; ΔH attributes degradation to environmental variability (H_f) or internal indistinguishability (H_b). This matters for coordination—agents unpredictable in their effects become unreliable partners (Dragan et al., 2013; Hadfield-Menell et al., 2016), H_b directly quantifies this failure mode.

4.5 Biological Precedent: Monitoring Statas and Actions in the Brain

Intelligence, as defined here, requires observing the (S, A, S') stream. A biological precedent exists in the mammalian thalamocortical loop, where thalamic nuclei receive copies of both sensory signals (S) and motor commands (A) via branching axons (Guillery, 2005). These are copies—not modulatory inputs—positioning the thalamus as an observer of the interaction, not a controller. Thalamic circuits operate on signal statistics—gain, synchrony, bandwidth—rather than semantic content (Sherman & Usrey, 2024; Cassidy et al., 2025). This suggests biology monitors interaction structure independently of task meaning. We do not claim the thalamus implements an IDT. Rather, it provides existence proof that copy-based observation of (S, A) streams can coexist with effective control—an architectural principle evolution discovered independently.

5 Results—Bi-Predictability Engineering Validation

We test whether current AI systems satisfy the operational conditions for agency and intelligence introduced above. This extends prior work showing that interaction information $MI(S, A; S')$ can flag behavioral anomalies in robotics and perception (Reid et al., 2025; Nazeri et al., 2025) by adding regime bounds and explicit criteria. We evaluate reinforcement-learning agents in continuous control and large language model agents in multi-turn interaction, computing P and ΔH from the (S, A, S') stream without access to model internals, reward shaping, or semantic content.

5.1 Bi-Predictability for Reinforcement Learning Agents (RL)

5.1.1 Experimental Setup

We evaluate continuous-control agents in MuJoCo (Todorov et al., 2012), trained with SAC and PPO (Haarnoja et al., 2018; Schulman et al., 2017) on HalfCheetah. Policies are frozen during evaluation. Metrics P, H_f, H_b, and ΔH are computed over fixed-length sliding windows; perturbations begin mid-evaluation after a baseline period. Results aggregate across seeds (11 SAC, 10 PPO). For threshold-based detection, seeds with unstable pre-perturbation baselines were excluded from detection-rate summaries (reported explicitly below), since calibration requires a stationary baseline.

5.1.2 Baseline Coupling

Under normal operation, Half-Cheetah exhibits $P = 0.33 \pm 0.02$ and $\Delta H = -0.56 \pm 0.22$, placing it below the classical ceiling and within the agentic regime. The negative ΔH indicates persistent asymmetry: backward ambiguity exceeds forward uncertainty, consistent with interventions that do not fully round-trip from outcomes back to internal causes. Table 3 contrasts this with the double pendulum baseline ($P \approx 0.48$, $\Delta H \approx 0$), separating physical and agentic regimes.

System	P	ΔH	Interpretation
Double pendulum	0.48	≈ 0	Physics: high coherence, symmetric prediction
Half-Cheetah (baseline)	0.33	-0.56	Agency: reduced coherence, asymmetric prediction

Table 3 *Bi-Predictability across physical and agentic systems. Agency reduces P and breaks predictive symmetry, confirming theoretical predictions.*

5.1.3 Drift Detection Coverage

We injected eight perturbation types spanning environment-side changes (e.g., forces/gravity) and agent-side degradation (e.g., observation/action noise). Across 168 perturbation trials, the IDT detected $89.3 \pm 15.1\%$ of perturbations, compared with $44.0 \pm 26.1\%$ using reward-based detection ($t = 7.95, p < 10^{-6}$). Individual components (P, ΔH, H_f, H_b) each detect several perturbations, and their union increases coverage because the signals respond to different failure modes.

5.1.4 Drift Detection Latency

IDT also detects degradation earlier. Median detection latency is 42 windows post-onset for IDT versus 184 for reward (Table 4), reflecting that reward integrates effects over many transitions whereas $P$and ΔHtrack coupling integrity at the transition level.

Metric	Median Latency (windows)
IDT	42
P	74
ΔH	67
H_f	69
H_b	75
Rewards	184

Table 4 *Detection latency (median windows after perturbation onset).*

5.1.5 What P Reveals That Reward Cannot

These results support the framework's central distinction between task performance and interaction quality. Baseline values ($P = 0.33$, $\Delta H = -0.56$) place the RL agent below the physical ceiling and show the predictive asymmetry expected in the agentic regime. The detection advantage (89% vs 44% coverage; $4.4\times$ lower median latency) follows from what the signals measure: reward integrates outcomes over many transitions, so coupling degradation often becomes visible only after failures accumulate. By contrast, P and ΔH track coupling at the transition level, so disruption is detectable immediately—even before returns degrade.

Because P and ΔH respond to different failure modes, their combination increases detection coverage beyond any single component (Fig. 4). Moreover, different perturbations produce distinct response patterns across P, H_f, H_b, and ΔH, suggesting a path toward attribution rather than a single undifferentiated alarm.

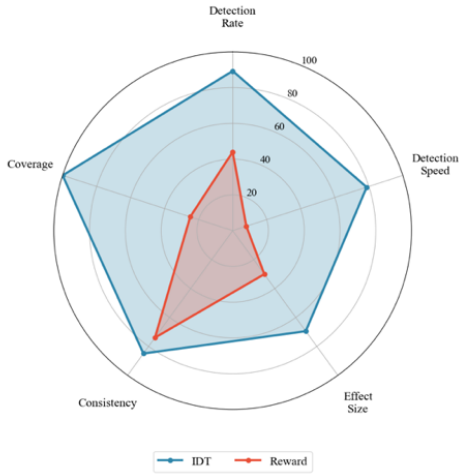

Figure 4 *DT diagnostic profile. IDT (blue) outperforms reward (red) across all five dimensions: detection rate, speed, effect size, consistency, and coverage.*

5.1.6 Meeting Agency and Intelligence Conditions

RL agents satisfy all three agency conditions: stochastic policies ensure choice ($H(A|S) > 0$), actions causally influence outcomes ($MI(A; S'|S) > 0$), and predictive asymmetry distinguishes them from passive physics ($\Delta H = -0.56 \neq 0$). They also satisfy learning—training maximizes $MI(S, A; S')$ towards cumulative reward. However, they lack self-monitoring and adaptation: no mechanism computes P from the agent's own (S, A, S') stream, nor can they adjust their sensors (S), effectors (A), or deployment environment (S'). By our definition, current RL agents exhibit agency and learning, but not intelligence (Table 5).

Condition		Criterion	Evidence	Achieved
Agency	Choice	$H(A\|S) > 0$	Stochastic policies (SAC, PPO)	Yes
	Effect	$MI(A;S'\|S) > 0$	Actions influence outcomes	Yes
	Asymmetry	$\|\Delta H\| > 0$	$\Delta H = -0.56 \pm 0.22$	Yes
Intelligence	Learning	$\uparrow MI(S,A;S')$ towards objective	Trained on (S,A,S',R) to maximize reward	Yes
	Self-monitoring	Computes P from own stream	No internal P computation	NO
	Adaptation	Adjusts $\{S\},\{A\},\{S'\}$	Spaces fixed by designers	NO

Table 5 *Agency and intelligence conditions (RL agents). RL agents satisfy agency (choice, effect, asymmetry) and learning, but lack self-monitoring—the defining gap between current AI and intelligence.*

5.1.7 Attribution Implications

P alone signals degradation but not direction. The decomposition into H_f and H_b provides diagnostic structure: H_f captures uncertainty about outcomes; H_b captures uncertainty about causes. Neither metric uniquely identifies the source—both agent and environment changes can affect either component. However, the pattern of responses narrows the search space, transforming blind troubleshooting into directed investigation. When H_f and H_b respond differently, the asymmetry localizes the breakdown. When both respond together, systemic changes are implicated. This diagnostic capacity—unavailable from reward or P alone—is what makes attribution actionable. Details of the attribution logic are formalized in Table 10 (Methods).

5.2 Large Language Model Drift Detection using P

5.2.1 Setup

To test generality beyond physical control, we evaluate Bi-Predictability in multi-turn dialogue. A student model (Llama 3.1 8B) interacts for 100–200 turns with three distinct teacher models (Claude, ChatGPT, Gemini) across 34 unique test–teacher–condition combinations (4,574 turns total). Conditions varied: normal (temperature 0.7, top_k 40) allowed unrestricted generation, while constrained (temperature 0.1, top_k 10) reduced response diversity, simulating capacity degradation.

Three baseline tests examined natural conversation dynamics using prompts designed to elicit varied questioning styles. Three perturbation tests evaluated sensitivity to conversational disruptions—contradictions, topic shifts, and non-sequiturs—injected at fixed intervals after a 30-turn baseline. We map dialogue into the (S, A, S') loop: S is accumulated context, A is the student response, and S' is the teacher's subsequent prompt. Metrics, H_f, H_b, and ΔH are computed from token-frequency distributions.

5.2.2 Quantifying Structural vs. Semantic Coherence in Large Language Models

We compare $P$and ΔHagainst two widely used baselines: embedding-based cosine similarity for structural consistency (Reimers & Gurevych, 2019) and LLM-as-a-judge for semantic quality (Zheng et al., 2023). Across conditions, P aligns strongly with structural consistency (significant correlation in 85% of cases; Table 6) but aligns less reliably with judge scores (44% of cases). This separation indicates that P primarily tracks interaction structure rather than semantic correctness—an interaction-quality signal that does not require embeddings or external evaluation models.

Metric	Correlation with Structure (Cosine Sim)	Correlation with Semantics (LLM Judge)
Prediction Efficiency (P)	**85%** (29/34 conditions)	44% (29/34 conditions)
Prediction Asymmetry (ΔH)	**76%** (26/34 conditions)	47% (26/34 conditions)

Table 6 *Relationship to structure and semantics. Across test conditions, Pand ΔH correlate more consistently with embedding-based structural similarity than with judge-based semantic scores, indicating that Bi-Predictability primarily tracks interaction structure.*

5.2.3 Perturbation Detection and Consistency Across Teachers

We inject three perturbation types (contradictions, topic shifts, non-sequiturs) at fixed turn positions. Using only token statistics, P and ΔH achieved 100% detection across all teacher models and perturbation types ($9/9$ trials per

condition, $p < 0.001$), matching the sensitivity of semantic judges (Cosine/GPT-4) but with significantly lower computational overhead. As shown in Fig. 5, deviations exhibit consistent signatures: P exhibits immediate instability at injection points—typically a sharp drop due to confusion or occasionally a spike due to fixation—while backward predictivity (H_b,) simultaneously increases. This confirms that structural coupling metrics are sufficient to flag semantic breakdowns without requiring heavy semantic evaluation

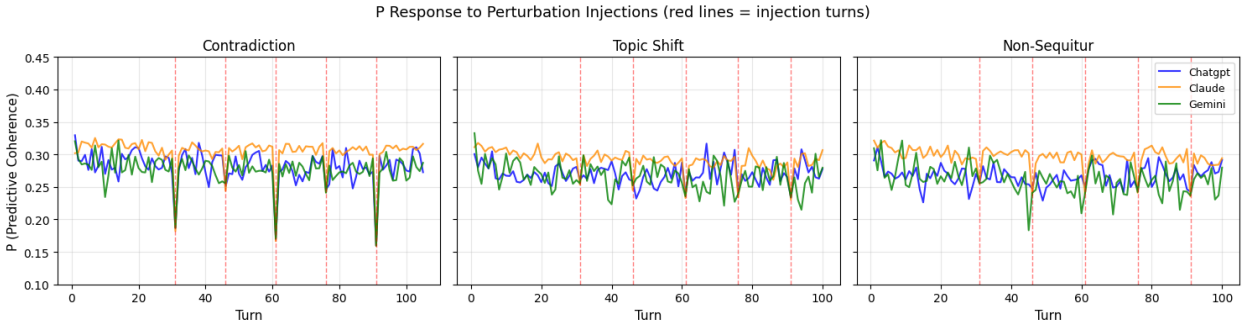

Figure 5 *Bi-Predictability under dialogue perturbations. P trajectory across multi-turn interactions with three teacher models under three perturbation types. Vertical markers indicate injection points; P drops sharply at each perturbation and typically recovers within 1–2 turns.*

5.2.4 Interpretation: Conditions for Agency and Intelligence

LLM agents satisfy the agency criteria at the interaction level: stochastic sampling provides choice; responses influence subsequent context; and ΔH indicates persistent predictive asymmetry. They also satisfy learning through next-token training. However, as summarized in Table 8, current LLM agents lack explicit self-monitoring and adaptation: they do not compute coupling quality nor can they reorganize their interface in response to degradation.

The IDT fills this gap. Unlike semantic evaluators such as cosine similarity or LLM judges—which introduce significant latency and model dependencies—the IDT operates directly on raw token statistics with negligible overhead. This computational efficiency allows it to transform the passive token stream into a real-time active control signal. By surfacing coherence deviations immediately, the IDT provides the necessary feedback to trigger the Coupled Agency Architecture's reflexive modulation—enabling the system to restore stability through context gating or parameter adjustment, rather than relying solely on fixed next-token probabilities

Condition		Criterion	Evidence	Achieved
Agency	Choice	$H(A\|S) > 0$	Stochastic sampling (temperature > 0)	Yes
	Effect	$MI(A;S'\|S) > 0$	Responses influence subsequent context	Yes
	Asymmetry	$\|\Delta H\| > 0$	$\Delta H < 0$ across all conditions	Yes
Intelligence	Learning	$\uparrow MI(S,A;S')$ towards objective	Trained on token sequences to predict next token	Yes
	Self-monitoring	Computes P from own stream	No internal P computation	NO
	Adaptation	Adjusts $\{S\},\{A\},\{S'\}$	Vocabulary and generation parameters (context window, top_p, top_k, max response) fixed by designers/users	NO

Table 7 *Agency and intelligence conditions in LLM agents. LLM agents satisfy agency and learning, but lack explicit self-monitoring and adaptation under our definition.*

6 Method

Method section is removed from this version.

7 Discussion

Currently, AI development focuses on scaling the internal model (learning). Our results suggest that reliable agency requires a parallel focus on the Information Architecture: the structural capacity to regulate coupling quality. By identifying Predicative Coherence (P) as the order parameter of interaction, we distinguish effective agency from mere throughput. The systematic reduction of P when actions are introduced reflects the informational cost of freedom; intelligence is not the elimination of this cost, but the active management of it via self-monitoring.

Within this framework, agency is the introduction of choice into the agent–environment loop, while intelligence requires learning plus explicit self-monitoring and adaptation. Actions add internal degrees of freedom that typically reduce raw predictability; managing this trade—rather than eliminating it—is the defining challenge of adaptive behavior. Both reinforcement-learning and large language model agents satisfy agency (choice, effect, asymmetry) and learning (increasing interaction predictability toward objectives). Yet neither satisfies self-monitoring nor adaptation: no current AI computes its own decision effectiveness from its own interaction stream, and state–action–outcome spaces remain designer-defined. Thus, under our operational definition, current AI exhibits agency and learning, but not intelligence.

Accordingly, there is a need for a metric that captures the "first-person" structural state of the agent, distinct from its third-person objective performance. While reward functions track external success, P quantifies the agent's "grip" on the environment—the bidirectional constraint where perception reliably dictates outcomes (forward predictability) and outcomes unambiguously reveal authorship (backward predictability). In biological systems, the independent failure of these constraints corresponds to distinct breakdowns requiring distinct recoveries. High forward uncertainty (H_f) means the world is opaque to the agent—outcomes remain unpredictable despite action. High backward uncertainty (H_b) means the agent is opaque to the world—different actions produce indistinguishable outcomes, as if the environment cannot read the agent's intent. Without this differentiation, an agent knows only that P dropped, not whether to adjust its predictions (H_f) or its legibility (H_b). Attribution is not diagnostic luxury—it is prerequisite for effective adaptation. Current AI systems are blind to these structural shifts; they pursue objectives even as causal coupling disintegrates. $P\ and\ \Delta H$ together provide the missing first-person metric: P measures coupling integrity, ΔH indicates where it fails.

The metric $P = MI(S, A; S') \,/\, H\_total$ operationalizes this by quantifying the fraction of total system entropy captured by the state-action-next-state coupling. Any significant deviation from baseline — regardless of direction — indicates the learned information structure no longer holds. The Information Digital Twin (IDT) monitors this coupling in real-time, supplying the regulatory layer missing from reward-based systems. By separating 'Task Performance' (the Agent) from 'Coupling Stability' (the IDT), the proposed Coupled Agency Architecture resolves the fragility of open-loop control. We identify Reflexive Modulation — the ability to gate observation and action bandwidths in response to statistical drift — as the critical mechanism for recovery. This mirrors the mammalian thalamus, which regulates signal transmission based on statistical properties rather than semantic content. While we define the information-theoretic specifications for these modulation interfaces, the specific control laws mapping coherence deviations to bandwidth adjustments remain a domain-specific engineering challenge for future work.

Collectively, these results establish that scalable intelligence depends not only on objective functions, but on explicitly engineered information coupling architectures—a structural layer that biological systems embody and current artificial systems must now adopt.

8 References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
2. Ashby, W. R. An Introduction to Cybernetics (1956)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.
4. Cassidy, R. M., Macias, A. V., Lagos, W. N., Ugorji, C., & Callaway, E. M. (2025). Complementary organization of mouse driver and modulator cortico-thalamo-cortical circuits. Journal of Neuroscience, 45(5).
5. Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley-Interscience.
6. D'Amour, A., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226), 1-61.
7. Dragan, A. D., Lee, K. C., & Srinivasa, S. S. (2013, March). Legibility and predictability of robot motion. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 301-308). IEEE.
8. Friston, K. (2010). The free-energy principle: a unified brain theory?. Nature reviews neuroscience, 11(2), 127-138.
9. Guillery, R. W. (2005). Anatomical pathways that link perception and action. Progress in brain research, 149, 235-256.
10. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (pp. 1861-1870). Pmlr.
11. Hadfield-Menell, D., Russell, S. J., Abbeel, P., & Dragan, A. (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems*, *29*.
12. Hafez, W. (2022, August). Information as Entanglement—A Framework for Artificial General Intelligence. In International Conference on Artificial General Intelligence (pp. 20-29). Cham: Springer International Publishing.
13. Haller, G. (2001). "Distinguished material surfaces and coherent structures in three-dimensional fluid flows." Physica D: Nonlinear Phenomena. (Signals that your comparison metric, FTLE, is based on the gold-standard definition.)
14. Horodecki, R., Horodecki, P., Horodecki, M., & Horodecki, K. (2009). Quantum entanglement. Reviews of modern physics, 81(2), 865-942.
15. Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2005, September). Empowerment: A universal agent-centric measure of control. In 2005 ieee congress on evolutionary computation (Vol. 1, pp. 128-135). IEEE.
16. Lakshminarasimhan, K. J., Xie, M., Cohen, J. D., Sauerbrei, B. A., Hantman, A. W., Litwin-Kumar, A., & Escola, S. (2024). Specific connectivity optimizes learning in thalamocortical loops. Cell reports, 43(4).
17. Nazeri, A., & Hafez, W. (2025). Entropy-Based Non-Invasive Reliability Monitoring of Convolutional Neural Networks. arXiv preprint arXiv:2508.21715.
18. Nielsen, M. A., & Chuang, I. L. (2010). Quantum Computation and Quantum Information. Cambridge University Press.
19. Press, W. H., et al. (2007). Numerical Recipes: The Art of Scientific Computing. Cambridge University Press. (For the ode45/Runge-Kutta integration method).
20. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., & Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. Journal of machine learning research, 22(268), 1-8.
21. Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2(1), 79-87.
22. Rawlings, J. B., Mayne, D. Q., & Diehl, M. (2020). Model predictive control: theory, computation, and design (Vol. 2). Madison, WI: Nob Hill Publishing.
23. Reid, C., Hafez, W., & Nazeri, A. (2025). Mutual Information Tracks Policy Coherence in Reinforcement Learning. arXiv preprint arXiv:2509.10423.

24. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084.
25. Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In Proc. of the international conference on simulation of adaptive behavior: From animals to animats (pp. 222-227).
26. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
27. Shannon, C. E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal, 27(3), 379–423
28. Sherman, S. M., & Usrey, W. M. (2024). Transthalamic pathways for cortical function. Journal of Neuroscience, 44(35).
29. Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). The curse of recursion: Training on generated data makes models forget. Nature, 632, 755–759.
30. Strogatz, S. H. (2018). Nonlinear Dynamics and Chaos. CRC Press. (Signals that you are using the standard physics definition of the system.)
31. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). The MIT Press.
32. Todorov, E., Erez, T., & Tassa, Y. (2012, October). Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems* (pp. 5026-5033). IEEE.
33. Usrey, W. M., & Sherman, S. M. (2019). "Transthalamic Pathways for Cortical Function." Journal of Neuroscience.
34. Wiener, N. (2019). Cybernetics or Control and Communication in the Animal and the Machine. MIT press.
35. Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.