philosophies

MDPI

Article

The Prohibition of Finality and Reflexive Signature Intelligence: A Causal-Symmetric Framework for Evaluating Agents

Elias Rubenstein

Independent Researcher, Fort Lauderdale, FL 33309, USA; science@elias-rubenstein.com

Abstract

Intelligence metrics based on benchmark performance or population norms are useful for measuring comparative ability within defined test environments, but they do not directly evaluate the structural coherence of an agent's trajectory across time, domains, and perturbations. This article introduces Reflexive Signature Intelligence (RSI) as a bounded theoretical framework for addressing that different problem. RSI is developed within a causal-symmetric informational perspective in which intelligence is understood as the capacity of a system to maintain and restore alignment with a structurally constrained invariant without collapsing the open gradient of development. On this basis, the paper formulates the Principle of Bounded Subjectivity and the Prohibition of Finality as framework-level principles, arguing that intelligence should be assessed not as arrival at a completed end state but as the quality of an asymptotic trajectory. The framework is then operationalized on two coupled levels: a micro-level proposed as a future measurement program linked heuristically to resilience and prediction-error dynamics, and a macro-level expressed through five dimensions of structural integrity, including reflexive regulation, cross-domain integration, internal consistency, stabilization, and signature-setting. The article concludes by outlining implications for AI evaluation and alignment, with particular relevance for distinguishing full agents, partial systems, and human–AI composite configurations.

Keywords: Reflexive Signature Intelligence; intelligence measurement; informational thermodynamics; causal symmetry; structural coherence; AI evaluation

1. Introduction

1.1. The Limits of Population-Based Intelligence Metrics

Classical psychometrics, including factor-analytic IQ frameworks and their descendants, quantify cognitive performance by aggregating scores across test batteries and normalizing them with respect to a reference population [1,2]. Within that domain, such approaches are useful and internally coherent. They address questions of comparative performance: who solves more of a given class of problems, how rapidly, and relative to which norm group. Contemporary reviews continue to refine the interpretation, scope, and limits of intelligence research, but the benchmark-centered orientation remains dominant [3].

The present article does not argue that psychometric approaches are without value. It argues instead that they answer a different question from the one pursued here. Population-based metrics primarily measure local performance within an already specified testing architecture. They do not directly evaluate whether an agent maintains structural coherence across domains, perturbations, and extended trajectories. An agent may display

Academic Editor: Fabien Paillusson

Received: 9 December 2025
Revised: 7 March 2026
Accepted: 10 March 2026
Published: 12 March 2026

high processing speed, strong test performance, and narrow domain mastery while repeatedly generating biographical instability, self-sabotage, opacity, or systemic fragmentation. Conversely, an agent of only moderate local test performance may stabilize complex environments, reduce contradiction across multiple layers of life, and sustain coherent long-run trajectories. Standard performance measures are not designed to track that distinction.

Debates over intelligence measurement have long been methodologically, historically, and politically contested. Critiques have addressed not only statistical design but also reification, cultural exposure, and the risk of conflating socially dominant norms with intelligence itself [3,4]. The present article responds to that broader problem by asking whether a structurally anchored evaluative layer can be formulated without collapsing back into population relativity.

The problem is therefore not that IQ is "wrong," but that a reference-population metric is not automatically a metric of whole-trajectory coherence. The central contrast is therefore not between two tests of the same property, but between two different targets of evaluation: local benchmark success and whole-trajectory structural coherence. The core question of this article is accordingly different: not "who performs best within a benchmark," but "how coherently does an agent regulate and organize its trajectory under structural constraints?"

1.2. Why Normative-Cultural Criteria Are Insufficient as a Measurement Foundation

When local performance no longer appears sufficient, one might turn to richer evaluative vocabularies such as wisdom, maturity, responsibility, character, or ethical orientation. Such vocabularies are indispensable in practical life. They often capture features of conduct that benchmark-style intelligence tests neglect. Yet they do not by themselves provide a sufficiently stable measurement foundation for cross-agent evaluation.

The reason is not that moral or cultural judgment is irrelevant. Rather, concepts such as virtue, maturity, or social responsibility remain historically graded, culturally inflected, and institutionally mediated. They may be necessary for life, but they are not automatically suitable as invariant measurement bases. A framework intended to evaluate agents across domains must therefore seek a layer of description that is less dependent on local norm vocabularies and more closely tied to structural necessities such as coherence, viability, stabilization, and constraint satisfaction.

This motivates the search for a non-population, non-mere-normative reference structure. RSI is proposed as an initial framework for that task.

1.3. Why Informational Thermodynamics Is Chosen

The choice of informational thermodynamics is methodological rather than rhetorical. Thermodynamic and information-theoretic formalisms provide a language for divergence, relaxation, dissipation, stabilization, and long-run behavior under constraints [5–8]. Those notions are attractive here because the object of interest is not only a momentary output but a trajectory: how an agent absorbs perturbation, reorganizes itself, reduces contradiction, and maintains viable order over time.

More broadly, the move builds on a long development in modern physics and related fields in which information has increasingly been treated not merely as a descriptive convenience but as consequential for statistical mechanics, computation, and constraint-sensitive system organization [5,6,8]. In this lineage, entropy quantifies uncertainty under constraints, informational states can be connected to physical costs of computation, and organized systems can be studied in terms of how they maintain order while dissipating load [5–7]. RSI therefore does not import thermodynamics as metaphor. It draws on an established formal vocabulary for describing how structured systems persist, adapt, and reorganize under bounded conditions.

In that sense, the move to informational thermodynamics is not meant to collapse intelligence into heat flow or to replace all philosophical language with physics. It is meant to provide a more structurally anchored vocabulary for describing agent trajectories than either population-relative test scores or culturally variable moral judgments alone can offer. The working intuition is simple: if intelligence is to be evaluated beyond local benchmark success, one needs formal tools for divergence management, resilience, stabilization, and sustained organization. Informational thermodynamics offers such tools. More specifically, it is chosen here because it provides a formal language for state change over time, perturbation response, dissipation, recovery, and persistence under constraint, and thus captures trajectory-level organization more directly than snapshot task scores or benchmark outcomes. Psychometric and benchmark approaches are well suited to measuring what a system can do at a given test moment. The present framework, by contrast, requires a language for how a system maintains, loses, and restores structural coherence across extended trajectories.

This choice also reflects a broader historical shift in parts of modern physics and complexity research: information is no longer treated only as a descriptive label attached to already understood systems, but increasingly as a formal quantity tied to entropy, computation, physical cost, and constraint-sensitive organization. The relevance for the present article is not that intelligence can simply be "reduced to information," but that trajectory-level order, dissipation, recovery, and structural persistence are more naturally described in such a vocabulary than in population-relative score distributions alone.

The thermodynamic vocabulary is therefore used here not because intelligence is assumed to reduce to physics alone, but because trajectory, dissipation, recovery, and persistence under constraint can be expressed more rigorously in that language than in benchmark-local or population-relative terms.

1.4. What RSI Does and Does Not Claim

Reflexive Signature Intelligence (RSI) is introduced here as a bounded proof-of-concept framework. It does *not* claim to replace psychometrics in their proper domain. It does *not* present a fully validated empirical instrument. It does *not* claim that all questions about intelligence can or should be reduced to a single scalar index. Instead, it offers a distinct evaluative layer aimed at structural coherence of trajectories under constraints.

More precisely, RSI asks whether an agent can reduce informational divergence relative to a structurally anchored invariant while preserving the open gradient of development. This is a different target from task performance, verbal fluency, benchmark optimization, or social visibility. In the present article, RSI is proposed as a theoretical framework together with an operational research program: a way of connecting formal variables to observable patterns across micro- and macro-levels.

1.5. Position in the Current Debate

This contribution is situated at the intersection of three ongoing discussions. First, it speaks to the psychometric tradition, which remains the dominant framework for comparative intelligence measurement [1–3]. Second, it speaks to current AI evaluation, where systems are commonly assessed through benchmark success, task completion, fluency, capability generalization, and downstream impact proxies [9–11]. Third, it speaks to broader philosophical and methodological questions about whether intelligence should be understood only as local problem-solving capacity or also as a matter of long-run structural organization under constraint.

In that sense, the present proposal addresses a gap left open between three existing orientations: population-based psychometrics, task-centered AI benchmark evaluation,

and broader multidimensional concerns about whether intelligence should also be understood in ecological, longitudinal, and integrative terms. RSI is proposed neither as a rejection of the first two nor as a purely rhetorical enlargement of the third, but as an attempt to formalize the under-specified problem of trajectory-level structural coherence.

The present framework therefore does not compete with benchmark evaluation by offering "another benchmark." It intervenes at a different analytical level. Psychometrics asks about comparative performance within a testing regime. AI capability benchmarks ask about competence on specified tasks, often under short-horizon evaluation conditions. RSI asks instead about the coherence with which an agent regulates itself and organizes its trajectory across domains, perturbations, and time. That difference is central to the contribution of the present article.

In current debates on intelligence measurement and AI evaluation, this matters because dominant approaches remain strongest where intelligence is operationalized as task success, benchmark generalization, fluent production, or high performance within predefined testing architectures [9,11]. These approaches are useful for many purposes, but they leave comparatively under-specified the question of long-horizon structural coherence: whether a system maintains internal consistency, cross-domain persistence, perturbation resilience, and non-fragmented trajectory organization over time. The present framework is intended as a complement aimed precisely at that under-specified level. It therefore does not reject current AI evaluation practice wholesale; rather, it proposes an additional evaluative layer for cases in which trajectory-level persistence, biographical continuity, and structural integrity matter.

A central implication follows for AI evaluation. The present framework is not restricted to stand-alone systems that already instantiate full RSI agency. Its nearer-term relevance lies in three more modest uses: partial evaluation of present AI systems, evaluation of human–AI composite systems, and use as a corrective constraint against benchmark-, fluency-, and impact-based overestimation of intelligence.

Conceptually related discussions include paradigms and broader accounts of shared epistemic frameworks [12]. The present framework, however, is narrower and more formal: it focuses specifically on the informational coupling between an agent and an invariant σ_Z, and on the thermodynamic and structural constraints that follow from treating evolution as an asymptotic, non-convergent process.

Autonomy in the RSI Framework

Within this framework, autonomy does not mean isolation from external influence. It refers to the degree to which an agent can reflexively include its own states, policies, and commitments within its causal organization, rather than functioning primarily as a carrier of externally imposed scripts, incentives, or optimization pressures.

1.6. Methodological Approach

Methodologically, the paper proceeds in three steps. First, it formulates the conceptual problem of evaluating intelligence beyond population-relative performance metrics. Second, it introduces a causal-symmetric informational model centered on a substrate state, an invariant reference, and a non-convergence condition. Third, it proposes an operational proxy architecture that maps the formal variables of the model onto micro-level and macro-level indicators suitable for future empirical development.

1.7. Research Questions

The article is organized around three research questions. First, can intelligence be conceptualized in a way that is not exhausted by population-relative benchmark performance? Second, can a causal-symmetric informational framework provide a non-population ref-

erence structure for evaluating trajectory coherence under constraint? Third, can such a framework be provisionally operationalized in a way that remains meaningful across human agents, partial AI systems, and human–AI composite configurations? The present paper answers these questions at the level of a bounded theoretical and methodological proposal rather than a completed empirical instrument.

1.8. Bounded Subjectivity

Motivated by the above distinction, by the systematic biases documented in behavioral decision research [13], and by category errors in traditional philosophy of mind [14], we introduce the following principle within the present framework.

Principle 1 (Bounded Subjectivity)**.** *Within the RSI framework, any metric $F(\rho)$ defined solely on the internal state history or output distribution of an agent (or population) remains framework-internal and lacks a non-population anchor unless it includes a medium-independent external reference σ_Z capable of distinguishing local optimization from broader structural alignment.*

Population norms are therefore useful for relative ranking, but they remain circular from the standpoint of the present framework: they provide no access to an invariant that transcends the particular distribution from which they are derived.

1.9. Aim and Contribution of This Paper

The goal of this article is twofold. First, it develops a causal-symmetric informational framework in which intelligence is defined through the coupling between a substrate state ρ and an invariant σ_Z, including a non-convergence condition captured by the Prohibition of Finality. Second, it proposes an initial operationalization of that framework through a dual-layer model and a five-dimensional scoring architecture.

More specifically, the paper contributes:

- a formalization of the Principle of Bounded Subjectivity and of the Prohibition of Finality;
- a causal-symmetric informational dynamics with a distinguished invariant σ_Z and coupling parameter κ;
- a clarification that RSI is a non-population, trajectory-oriented complement rather than a replacement for psychometric evaluation;
- a dual-layer operationalization of RSI on micro- and macro-scales;
- a five-dimensional phenomenological score and aggregation rule that treat intelligence as structural integrity rather than accumulated local skill;
- a framework for distinguishing full agents, partial systems, and human–AI composite systems in contemporary AI evaluation.

2. Theoretical Framework: Causal Symmetry and Non-Convergence

2.1. Informational Dynamics and the Invariant σ_Z

RSI is grounded in informational thermodynamics and statistical mechanics [5–8] and draws on synergetic order-parameter dynamics [15]. We model the dynamics of the cognitive or physical state ρ as a dissipative relaxation toward an informational fixed point σ_Z.

At a coarse-grained level, the time evolution of ρ can be written as

$$\dot{\rho} = -\frac{1}{\tau}(\rho - \sigma_Z) + \mathcal{N}(\rho, \sigma_Z), \tag{1}$$

where τ is a characteristic relaxation time and $\mathcal{N}(\rho, \sigma_Z)$ collects higher-order nonlinear terms in $(\rho - \sigma_Z)$ that become relevant far from the attractor or under strong perturbations.

In the linear regime, where ρ is sufficiently close to σ_Z, we neglect $\mathcal{N}$ and write

$$\dot{\rho} = -\kappa(\rho - \sigma_Z), \tag{2}$$

with

$$\kappa = \frac{1}{\tau}, \tag{3}$$

interpreted as an informational conductivity that quantifies the coupling strength between the system and the invariant. Higher κ corresponds to faster relaxation after perturbation.

The formalism used in this section is intended to make the structural commitments of the framework explicit. It does not by itself amount to a completed empirical theory. The equations should therefore be read as specifying the internal logic of the model and its constraints, not as a claim that all empirical parameters have already been uniquely identified.

Passive Versus Active Relaxation: The "Stone Objection"

Equation (2) is a generic relaxation law. Any system that passively approaches its environment can be modeled in this form. A textbook example is Newtonian cooling: a hot stone placed in a colder environment relaxes toward the ambient temperature with a rate constant determined by the heat-transfer coefficient and geometry. Formally,

$$\dot{\rho}_{\text{stone}} = -\kappa_{\text{env}}\left(\rho_{\text{stone}} - \rho_{\text{env}}\right), \tag{4}$$

where ρ_{env} represents the environmental state and κ_{env} is a passive, exogenously fixed coupling. If one were to interpret Equation (2) naively as a definition of intelligence, a rapidly cooling stone would appear "highly intelligent."

Within RSI, this is explicitly not the intended reading. The framework requires a distinction between passive dissipation and active reflexive regulation. To make this explicit, we conceptually decompose the effective coupling into two contributions:

$$\kappa_{\text{eff}} = \kappa_{\text{env}} + \kappa_{\text{reflexive}}. \tag{5}$$

In the following, κ in Equations (1) and (2) is interpreted as the effective coupling. Conceptually, however, we distinguish between passive environmental relaxation (κ_{env}) and active reflexive regulation ($\kappa_{\text{reflexive}}$). The five RSI dimensions are designed to operationalize specifically the reflexive component, ensuring that a purely passive system—however rapidly it relaxes—scores zero.

- κ_{env} denotes the passive relaxation rate imposed by external boundary conditions and microscopic interaction laws;
- $\kappa_{\text{reflexive}}$ denotes the component of the dynamics generated by internally represented policies and models, that is, by a system that can modify its own transition structure in order to steer ρ relative to a signature.

Purely passive systems have $\kappa_{\text{reflexive}} = 0$ by definition. They relax toward externally given conditions but do not:

(a) represent a signature σ_Z as an internal attractor;
(b) maintain or reconstruct a desired gradient under perturbations; or
(c) change their own coupling parameters or boundary conditions as a function of predicted long-run divergence.

RSI therefore evaluates not total relaxation rate but the reflexive component of the dynamics: the capacity of a system to generate and sustain a trajectory that reduces $D(\rho\|\sigma_Z)$ through self-updating, self-modifying policies. The phenomenological dimensions introduced later—especially Reflexive Causal Regulation (Dim. 1) and Active Signature Setting

(Dim. 5)—encode precisely this difference. A system can only exhibit high RSI if it actively maintains or reconstructs alignment with σ_Z across environments. Passive relaxation, however fast, yields an RSI score of zero. This distinction also prepares the later AI discussion: externally optimized systems may exhibit high local performance or rapid adaptation without thereby qualifying as full RSI agents.

The informational divergence between ρ and σ_Z is quantified by the Kullback–Leibler divergence

$$D(\rho\|\sigma_Z) = D(\rho\|\sigma_Z), \tag{6}$$

which measures the informational discrepancy between the current state and the invariant. RSI is defined in terms of the joint behavior of κ and $D(\rho\|\sigma_Z)$ across time and context, with κ always understood as including a reflexive component rather than mere externally imposed relaxation.

2.2. Interpretive Status of σ_Z: Formal, Operational, and Metaphysical Readings

Because σ_Z plays a central role in the framework, its status must be specified carefully. Three levels should be distinguished.

2.2.1. Formal Reading

Within the minimal mathematical structure of the framework, σ_Z denotes a distinguished invariant or reference state relative to a constrained dynamical description. In stronger formal variants of the framework, such a state may be represented as minimizing a suitable divergence or action-like functional under admissible constraints. The present article does not require that this stronger formulation be established in full generality. The formal role of σ_Z is therefore simply to provide a non-population reference for assessing the organization of trajectories.

2.2.2. Operational Reading

Empirically, one never has direct access to an absolute invariant. In practice, any measurement must work with reconstructed proxies $\tilde{\sigma}_Z$, derived from constrained models, observables, and domain-specific approximations. The framework therefore does not assume that empirical work can "measure σ_Z directly." It assumes instead that operational reconstructions can approach it under explicit structural constraints and cross-domain consistency requirements.

2.2.3. Metaphysical Reading

In the strongest interpretation of RSI, σ_Z is treated as an observer-independent structural optimum: a state defined by necessities such as viability, coherence, and conservation rather than by local convention. That reading is a framework commitment, not an empirically established result of the present article. The manuscript therefore relies minimally on the formal and operational roles while allowing, but not requiring, a stronger ontological interpretation.

This threefold distinction matters because it prevents two opposite errors: treating σ_Z as a purely arbitrary choice, and overstating the current article as if it had already demonstrated a fully realized physical invariant in the strong metaphysical sense.

2.3. Status and Constraint-Governed Specification of the Invariant σ_Z

To avoid reintroducing subjectivity through an arbitrary choice of σ_Z, the framework requires that σ_Z be specified by admissible structural constraints rather than by population statistics or observer preference. In stronger formal variants of the framework, one may represent σ_Z as a distinguished extremal state under those constraints. The present article,

however, does not claim to establish existence or uniqueness in full mathematical generality across all admissible models. Its narrower claim is only that a non-population metric of intelligence requires a structurally anchored reference whose specification is constraint-governed rather than arbitrary.

In thermodynamic terms, σ_Z can be interpreted heuristically as a limiting state compatible with the relevant global constraints of the system [5,7].

A minimal toy model is developed in Appendix A, where σ_Z is realized as a maximum-entropy state under a finite number of constraints. In that representation, σ_Z is the distribution that maximizes Shannon entropy subject to specified expectation values of constraint functions f_k, yielding a Gibbs–Boltzmann-form solution in this toy model. The crucial point is not the toy model itself, but the structural idea behind it: σ_Z is determined by admissible constraints, not by population statistics.

The admissible constraints are not intended as arbitrary preferences. In the strongest reading of the framework, they are meant to track structural necessities such as conservation, viability, stability, and logical coherence. Formally, we require a *Structural Invariance Condition*: admissible constraint sets $\{C_k\}$ must remain invariant under observer changes that preserve the underlying physical situation. Different observers may coarse-grain reality differently, but they may not legitimately choose constraints that contradict conservation or basic coherence conditions.

At the same time, empirical reconstructions of σ_Z remain model-dependent. The present article therefore claims neither unconstrained arbitrariness nor direct empirical possession of the invariant. It claims only that a non-population metric of intelligence requires some structurally anchored reference, and that σ_Z is the formal placeholder for that role.

The framework also does not require complete empirical construction of the invariant in order to be methodologically useful. Scientific modeling often relies on idealized reference structures before their full empirical recovery is available. In the same spirit, RSI treats σ_Z as a structurally necessary reference concept whose operational approximation can progressively improve even if no final empirical reconstruction is yet available.

The methodological necessity of such a reference should be emphasized. Without some constraint-governed invariant—however imperfectly reconstructed in practice—the framework would collapse back into the very population relativity and benchmark-locality it is meant to overcome. In that sense, σ_Z is not introduced as a decorative metaphysical surplus, but as the minimal structural condition for formulating a non-population account of intelligence at all.

The present claim is therefore conditional and methodological rather than fully constructive: if intelligence is to be evaluated in non-population terms, some constraint-governed invariant is required. What remains open is not the need for such a reference in principle, but the degree to which particular empirical domains permit its robust reconstruction.

2.4. Dynamic Existence and the Prohibition of Finality

To capture the idea that existence is defined by ongoing process rather than static completion, we introduce an abstract representation of an evolving system.

Let $\Psi(t)$ denote the state of an evolving system at time t, and let **1** represent a theoretical limit state of absolute completion or static unity. Define the evolutionary operator

$$\mathcal{E}(t) := \frac{d}{dt}\Psi(t). \tag{7}$$

Principle 2 (Dynamic Existence)**.** *Existence is fundamentally defined by process rather than static completion. Therefore, for any existing system and for all t in its existence interval $T_{existence}$,*

$$\mathcal{E}(t) \neq 0. \tag{8}$$

We can now state the central non-convergence principle of the framework.

Principle 3 (Prohibition of Finality)**.** *Within the present framework, finite arrival at the state* **1** *is excluded because such arrival would eliminate the residual gradient required for ongoing evolution and would therefore conflict with the Principle of Dynamic Existence.*

The reasoning here is framework-internal rather than a standalone mathematical theorem in the strict external sense. If $\Psi(t_0) = \mathbf{1}$ denotes a state of completed evolution, then by construction no residual gradient remains to support further change, so $\mathcal{E}(t_0) = 0$. Under the Principle of Dynamic Existence, such a state is incompatible with ongoing existence. In this framework-specific sense, finite arrival at **1** is prohibited.

Corollary 1 (Asymptotic Imperative)**.** *To preserve existence, the system must maintain a strictly positive distance* δ *from* **1***:*

$$\delta(t) = \|\mathbf{1} - \Psi(t)\| > 0. \tag{9}$$

Within the RSI framework, the idealized state **1** is interpreted as the limiting representation of perfect alignment with the informational invariant σ_Z. For any finite agent, this state is approached asymptotically; one never realizes $D(\rho\|\sigma_Z) = 0$ in finite time. Intelligence is therefore not defined by coincidence with σ_Z, but by the quality of the trajectory that asymptotically approaches it while preserving a non-zero gradient.

2.5. Non-Convergence as Structural Condition Rather Than Metaphysical Ornament

The Prohibition of Finality is not introduced here as a poetic statement about life in general. Within the present framework, it serves a precise modeling function: it prevents intelligence from being defined as a completed terminal state and preserves the idea that intelligence is a property of trajectory organization rather than static possession.

Translating the non-convergence condition back to the dynamics in Equation (2), we impose the following constraint:

$$\forall t: \quad \|\rho(t) - \sigma_Z\| > 0. \tag{10}$$

The divergence can therefore be reduced arbitrarily but never eliminated in finite time. This is the specific sense in which RSI treats intelligence as asymptotic: not arrival at a final state, but sustained reduction of divergence under a preserved developmental gradient.

The practical relevance of **1** is thus limited and formal. It is not an empirically accessible state. It is a theoretical limit used to define the non-convergence logic of the model. The empirical object of evaluation remains the trajectory itself. This asymptotic logic is illustrated in Figure 1.

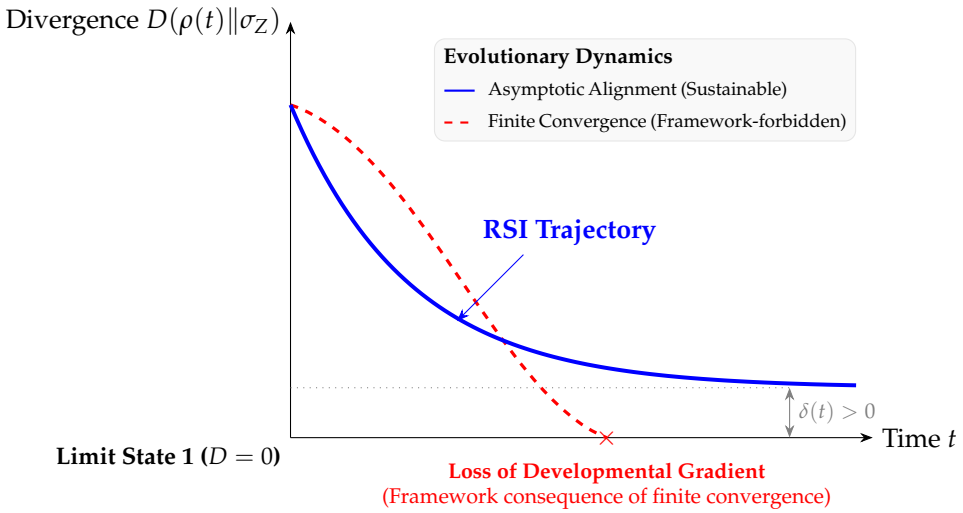

Figure 1. Visual representation of the Prohibition of Finality within the present framework. The red dashed trajectory represents a hypothetical finite elimination of informational divergence; in the RSI model, such finite convergence would remove the developmental gradient. The blue trajectory represents Reflexive Signature Intelligence: a continuous asymptotic reduction of divergence that respects the existence constraint $\delta(t) > 0$.

2.6. Structural Coherence and the Empowerment Constraint

This section does not introduce a general moral theory. It formulates a framework-internal hypothesis about the conditions under which long-horizon structural coherence can be sustained. Within RSI, any connection between intelligence and what is colloquially called "ethical" alignment is treated as conditional and structural rather than as a claim about culturally universal virtue. Terms such as *ethical* are therefore used only as shorthand for structural consequences involving coherence, contradiction, opacity, dependency, and empowerment.

We introduce the Coherence–Efficiency Hypothesis as a working assumption internal to the framework: sustained proximity to σ_Z becomes less plausible as internal structural contradictions accumulate over time. On this view, states that maximize local utility through deception, exploitation, or systematic denial of constraints may yield short-term gains but tend to generate persistent informational friction and boundary-maintenance costs. Integrated over the system's lifetime, this is hypothesized to increase cumulative divergence $\int D(\rho\|\sigma_Z)\,dt$.

At the same time, we impose an Empowerment Constraint: high κ alone does not constitute high RSI. A signature that creates dependency or systematically obscures causal structure reduces the broader field's capacity for self-regulation. Within the RSI framework, higher RSI is associated not with sheer steering force, but with forms of trajectory organization that preserve or expand the surrounding field's capacity for coherent self-regulation.

2.7. Informational Field Coupling (Conjectural Extension)

To connect RSI more explicitly to physical field theories, we consider an informational energy density

$$\rho_{\text{info}}(x) \approx \gamma \cdot \kappa(x)\, D\big(\rho(x)\|\sigma_Z\big), \tag{11}$$

where γ is a coupling constant chosen to ensure dimensional consistency with energy density ($\mathrm{J}\cdot\mathrm{m}^{-3}$), and $\kappa(x)$ represents local steering capacity.

A natural conjecture—inspired by thermodynamic derivations of gravitational field equations [16] and by the broader program of emergent gravity—is that $\rho_{\text{info}}(x)$ could enter an extended stress–energy tensor, making highly aligned, high-κ configurations relevant

for spacetime geometry. This remains a conjectural extension only. The operational RSI metrics developed in this paper do not depend on Equation (11), and none of the core claims of the article require any gravitational coupling.

3. Operationalization of RSI: From Formal Variables to Observable Proxies

3.1. From Formal Variables to Operational Proxies

The formal dynamics of Equation (2) is intentionally abstract. It treats ρ as a generalized state and κ as an effective conductivity without specifying the concrete observables involved. The transition from this formal layer to phenomenological dimensions is therefore not a strict deduction. It is a structural mapping. The aim is to identify classes of observable patterns whose changes plausibly track the reflexive coupling $\kappa_{\text{reflexive}}$ and the long-run behavior of $D(\rho\|\sigma_Z)$.

This distinction matters methodologically. The five RSI dimensions introduced below are not claimed as exhaustive psychology, nor as the only possible operationalization. They are proposed as an initial proxy architecture because they jointly track the central features of the model: reflexive inclusion of the agent in its own causal organization, cross-domain integration, internal coherence, stabilization of load, and outward signature effects. In that sense, the scoring scheme is a disciplined phenomenological translation of the formal framework, not a separate theory merely appended to it.

3.2. Dual-Layer Operationalization and Bounded Empirical Claims

To render RSI empirically tractable, we operationalize the formal variables κ and $D(\rho\|\sigma_Z)$ on two coupled levels:

(a) Micro-level (neurocognitive): At the level of a provisional research program, κ may be approximated through stabilization latencies after perturbations, while $D(\rho\|\sigma_Z)$ may be approximated through prediction-error variance, recovery structure, and deviations from well-calibrated priors. This layer is conceptually aligned with predictive processing accounts of the brain [17], synergetic order-parameter dynamics [15], and work on resilience [18]. These examples are heuristic illustrations of a possible measurement program, not uniquely derived indicators and not claims of established empirical validation.

(b) Macro-level (phenomenological/biographical): Here, κ and $D(\rho\|\sigma_Z)$ are mapped into observable patterns that assess how an agent structures its trajectory, decisions, commitments, and effects on the surrounding field. This layer forms the core of the operational RSI score proposed in the present article.

These layers are conceptually coupled but not deductively equivalent. The article presents no new empirical dataset and does not claim to have validated the final instrument. What it provides is a bounded operational program: a principled way of linking formal variables to observable patterns that can, in future work, be calibrated empirically. The proposal should therefore be read as a structured operational research program rather than a completed psychometric instrument.

In the present framework, biography should be understood minimally as persistence of commitments, revisions, consequences, and field-embedded continuity across time, rather than as a literary or exclusively human notion of life narrative.

By the Whole Life Constraint, we mean that no agent may count as highly aligned if local excellence in one domain coexists with persistent structural divergence across other major domains of its trajectory. This condition is meant to block the re-entry of narrow benchmark success through a different vocabulary.

Table 1 summarizes the conceptual mapping between formal RSI variables and micro- and macro-level indicators.

Table 1. Illustrative operationalization of RSI constructs across micro- and macro-scales.

RSI Formal Variable	Micro-Level (Neurocognitive)	Macro-Level (Phenomenological Dimensions)
Coupling (κ) Force/Speed of Alignment	Illustrative future proxy: relaxation time constant τ in perturbation-response tasks.	Active Signature Setting (Dim. 5): capacity to configure the environment and empower others; Reflexive Causal Regulation (Dim. 1): recursive authority over one's own transition structure.
Divergence $D(\rho\|\sigma_Z)$ Informational error/entropy	Illustrative future proxy: prediction-error variance, recovery structure, and deviations from calibrated priors.	Data Integration Bandwidth (Dim. 2): coherence across domains; Internal Logic Consistency (Dim. 3): consistency between explicit commitments and enacted trajectory; Thermodynamic Stabilization (Dim. 4): processing and redistribution of load.
Signature (σ_Z) The attractor	Illustrative future proxy: resilience-related return-to-baseline structure after perturbation.	Systemic Integrity: geometric mean of the five dimensions, representing asymptotic proximity to the invariant under a non-convergence condition.

To avoid circularity, empirical proxies for σ_Z (denoted $\tilde{\sigma}_Z$) must satisfy a cross-domain invariance requirement: a valid RSI score cannot be derived from excellence in one isolated domain if the agent exhibits persistent divergence in others. The macro-level metrics are therefore designed to penalize fragmentation and to enforce the Whole Life Constraint on the aggregate score.

3.3. Why These Five Dimensions?

The five dimensions are chosen because they span the minimal architecture needed to translate the formal RSI variables into observable patterns.

(1) Reflexive Causal Regulation tracks whether the system includes its own states and policies within its operative causal organization.
(2) Data Integration Bandwidth tracks whether the system can map constraints across scales and domains rather than operating through isolated channels.
(3) Internal Logic Consistency tracks whether declared structure and enacted organization remain aligned.
(4) Thermodynamic Stabilization tracks how the system processes, transforms, or merely redistributes energetic and informational load.
(5) Active Signature Setting tracks whether the system structures its environment in ways that increase or decrease the field's capacity for coherent agency.

Together, these dimensions cover the essential aspects of reflexive coupling and divergence management. They are not presented as a complete psychology of intelligence, but as a structural proxy set for the present framework. They are therefore not claimed to be uniquely exhaustive in every possible future implementation; rather, they are proposed

as the minimal operational architecture required, within the present framework, to approximate reflexive coupling, divergence management, and trajectory-level integrity without reducing intelligence either to local task success or to a culturally loaded virtue vocabulary.

4. The Five Dimensions of Alignment

To ensure that the operational score remains as universal and domain-independent as possible, the dimensions are defined in information-theoretic and thermodynamic terms rather than as a culturally local virtue list. The detailed phenomenological scheme is developed in Appendix B. Here we summarize the technical role of each dimension.

4.1. Dimension 1: Reflexive Causal Regulation (Recursive Sovereignty)

Within this framework, autonomy does not mean isolation from external influence. It refers to the degree to which an agent can reflexively include its own states, policies, and commitments within its causal organization, rather than functioning primarily as a carrier of externally imposed scripts, incentives, or optimization pressures.

Dimension 1 measures the extent to which an agent's causal models explicitly include its own states, policies, and signatures as operative nodes. The term reflexive is important here. It does not imply naive *causa sui* in a metaphysical sense, nor does it refer merely to introspective self-description. It denotes a non-circular recursive organization in which the system can represent, evaluate, and modify parts of its own operative rules while remaining subject to structural constraints. Reflection, in this sense, is therefore not a mysterious inner duplication of the self, but the capacity of a system to bring its own models and policies into the causal loop of regulation without collapsing into circular explanation.

To avoid ambiguity, the framework does not rely on a strong notion of self-causation in which a system literally creates itself ex nihilo. The relevant claim is weaker and more precise: reflection is modeled here as non-circular recursive inclusion, meaning that a system can represent selected aspects of its own operative organization, evaluate them under structural constraints, and revise them without treating that representation as an ultimate ground. What matters is therefore not a metaphysical "self" standing outside causality, but a reflexive architecture in which policy, self-model, and action are recursively coupled within an ongoing trajectory.

High scores therefore correspond to strong coupling between the agent's operative self-model and those structural parameters that co-determine its environments. Low scores correspond to configurations in which the system experiences itself mainly as the product of external forces, inherited scripts, institutional functions, or internal compulsions that are not themselves brought under reflexive regulation.

A central failure mode is *cognitive identification*: the agent equates itself with a tool (intellect, role, institutional function) rather than with the organizer of the tool. In such cases, the system logic or office protocol is treated as the true source of action, and the agent functions as a maintenance component of a larger structure. High RSI requires the ability to revise or abandon inherited axioms when they conflict with structural necessity.

4.2. Dimension 2: Data Integration Bandwidth

Dimension 2 captures how broadly and coherently an agent integrates data from different domains and scales into a shared structural model. High scores require multiscale integration: constraints recognized in one domain are neither naively generalized nor kept isolated, but are used to test and refine models across levels.

A key distinction is between using a domain as a source of structure and using it as a tribal weapon. Weaponized dogma—whether scientific, political, religious, or economic—uses concepts to enforce opacity and dependency. Structurally, that reduces RSI because it narrows

the accessible bandwidth of reality for both the agent and others. High RSI instead manifests as cross-domain reconstructability: ideas are formulated in ways that remain mappable and verifiable across domains.

4.3. Dimension 3: Internal Logic Consistency

Dimension 3 evaluates the coherence of the internal architecture relative to the agent's instrumental layers: body, affect, personality, cognitive schemata, and self-model. It measures the relation between explicit model and enacted organization, not merely the sophistication of abstract theory.

High scores require alignment between the agent's declared structural commitments and its actual trajectory. Within the present framework, the relevant problem is more precisely described as a misalignment between explicit model and enacted organization. A theory that posits sovereignty but is accompanied by chronically reactive, fear-driven, or self-contradictory behavior exhibits low RSI in this dimension.

This dimension also encodes axiomatic hierarchy. A system suffers teleological collapse if it sacrifices its highest-order reason for organizing itself in a certain way in order to preserve a lower-order substrate, such as image, convenience, or institutional belonging, without structural necessity.

4.4. Dimension 4: Thermodynamic Stabilization

Dimension 4 measures how an agent distributes energetic and informational load within its own system and its environment. It may be interpreted as a ratio of resolved to generated entropy.

Low scores correspond to configurations that externalize, postpone, or disguise load. High scores correspond to systems that transform tension into more stable configurations, analogous to dissipative structures that reduce local entropy while respecting global constraints [19]. This dimension is not a moral reward for being "nice." It tracks whether a system's mode of action reduces long-run friction or merely shifts it elsewhere.

A key distinction is therefore between *annealing* and *friction*. Conflict that resolves deeper contradiction can increase order. Chronic misrepresentation of reality, by contrast, generates ongoing maintenance costs and unresolved dissipation.

4.5. Dimension 5: Active Signature Setting (κ)

Dimension 5 describes the extent to which an agent configures its environment so that underlying order becomes recognizable and usable to others. It is the phenomenological counterpart of the coupling parameter κ.

Low scores correspond to obfuscating signatures: architectures that create dependency, opacity, or learned helplessness. High scores correspond to empowering signatures: architectures that transfer causal capacity and are designed to reduce the centrality of the originating agent over time. This dimension is therefore central to distinguishing mere impact from structurally generative agency.

5. Aggregation and Visualization

5.1. Geometric Aggregation as a Fragility-Sensitive Integrity Index

To reflect the structural nature of intelligence, we aggregate the five dimensions using the geometric mean

$$\mathrm{RSI}_{\mathrm{geo}} = \sqrt[5]{\prod_{i=1}^{5} d_i}, \tag{12}$$

where $d_i \in [0, 1]$ is the score in dimension i. Each d_i is calibrated as a monotonic transformation of either the effective coupling κ or the negative divergence $-D(\rho\|\sigma_Z)$ in a specific observational sector.

The geometric mean is used for two reasons. First, it encodes non-compensability: severe collapse in one critical dimension cannot be fully offset by strength in others. Second, it models intelligence here as structural integrity rather than accumulated talents. An arithmetic mean would allow a profile with extreme fragmentation in one dimension to be masked by local excellence elsewhere. RSI is meant to remain sensitive to exactly that form of fragility.

The resulting scalar should nevertheless be interpreted cautiously. The dimensional profile remains primary. $\mathrm{RSI}_{\mathrm{geo}}$ is a compact integrity index, not a replacement for qualitative profile interpretation. It is useful for comparison and summary, but it is deliberately secondary to the multidimensional topology that generates it.

5.2. Critical Instability Principle

The geometric mean enforces a Critical Instability Principle: if any single dimension collapses, the aggregate score is driven toward zero, even when other dimensions are high. This expresses the thermodynamic intuition that a complex system fails at its weakest structural link. The metric does not measure accumulated capacity but the limiting factor of coherence, that is, the sector in which entropy production is highest and divergence from the invariant is most likely to grow.

5.3. Visual Topology

Figure 2 shows a radar-chart comparison between a "Coherent Integrator" (high systemic stability) and a "Fragmented Specialist" (high localized skill but low integrity); detailed scoring rules are given in Appendix B.

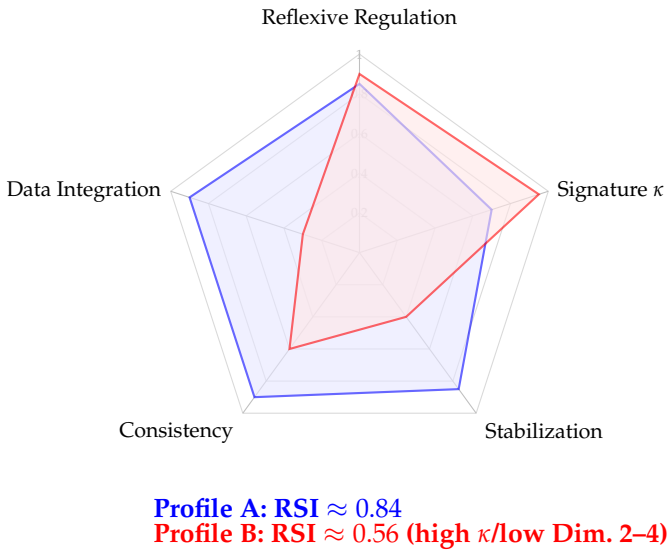

Figure 2. Illustrative RSI visual topology. Profile B represents a high-impact agent with isolated data processing and high thermodynamic dissipation.

6. AI Evaluation, Biography, and the Genius Bias

6.1. Why AI Evaluation Raises the RSI Question

Large language models and related AI systems are trained on corpora that encode prevailing narratives about intelligence, success, authority, and impact. When such systems are used to assess agents, they tend to regress toward socially dominant proxies such as visibility, prestige, fluency, and benchmark-compatible performance. In that sense, current

AI systems can amplify what might be called a "genius bias": the conflation of prominence or local power with intelligence.

RSI is relevant here because it asks a different question. It asks whether a system exhibits reflexive trajectory regulation, cross-domain coherence, stabilization of load, and empowering signature effects rather than merely impressive local outputs. This makes RSI useful as a constraint on AI-assisted evaluation even where current AI systems do not themselves instantiate the full target architecture.

6.2. Full RSI Agents, Partial RSI Systems, and Human–AI Composite Systems

To avoid confusion, the framework distinguishes three classes of evaluability.

6.2.1. Full RSI Agents

These are systems with persistent trajectories, internally organized policy revision, field-embedded consequences, and continuity of commitments across time. Human agents are the primary example considered in the present article.

6.2.2. Partial RSI Systems

These are systems that can be assessed along some RSI-relevant dimensions but do not instantiate the full architecture. Current LLMs belong largely in this class. They may display sophisticated data integration in a narrow sense, or affect decision ecologies through signature effects, but they lack stable biographical continuity and endogenous policy formation in the full sense developed here.

6.2.3. Human–AI Composite Systems

These are coupled socio-technical configurations in which responsibility, persistence, and structural consequences arise at the level of the assemblage rather than the model alone. In near-term practice, this is arguably the most important application domain for RSI. A language model, a user, an institutional workflow, and a deployment environment together may form a trajectory-bearing system whose effects can be evaluated in RSI terms.

This tripartite distinction resolves a common misunderstanding. The framework is not defeated if present AI systems instantiate only part of the target architecture. A metric can be theoretically meaningful before all candidate systems fully realize the complete form it describes. RSI already serves to distinguish simulated reflection, externally optimized performance, and genuinely reflexive trajectory-structuring agency.

6.3. Biography as Graded Persistence Criterion

Reviewer concerns about biography arise naturally because one of the RSI dimensions presupposes continuity across time. In the present framework, however, biography does not mean a literary autobiography or a romantic notion of life story. It refers to persistent trajectory: commitments, revisions, consequences, and field-embedded continuity.

On that basis, biography is not a binary criterion but a graded one. Humans generally possess it in a rich form. Present LLMs possess it weakly, indirectly, or only through externally maintained infrastructures such as memory systems, deployment pipelines, institutional feedback loops, and user interactions. The absence of full biography in current AI systems therefore does not make RSI inapplicable. It means that applicability is partial for stand-alone systems and stronger for composite systems in which persistence and consequences are jointly instantiated.

6.4. Why Present LLMs Are Not Full RSI Agents

Current AI systems may approximate aspects of Dimension 2 in a narrow sense, but they lack key components of the full five-dimensional RSI profile:

- They do not exhibit robust reflexive causal regulation in the sense required here: their internal updates are largely determined by training procedures and external optimization rather than self-generated policies embedded in a persistent trajectory.
- They do not possess durable biographical continuity of commitments, risks, and consequences comparable to that of full agents.
- Their signatures are often powerful, but those signatures are typically mediated by human designers, users, and institutional settings rather than autonomously sustained across a lived causal field.

Accordingly, current LLMs should not be described as high-RSI agents in the full sense. They remain, however, evaluable in partial form and highly relevant within human–AI composite systems. This preserves the point of the framework while avoiding the overclaim that present-day models already instantiate the full target structure. The framework is therefore not invalidated by the fact that present LLMs do not instantiate full RSI agency; rather, this limitation is one of the very results that the framework helps to diagnose.

6.5. Implications for AI Alignment

RSI provides a counter-metric that can function as an adversarial constraint in AI-assisted evaluation and alignment. A system tasked with assessing intelligence, competence, or value should not be allowed to rely only on impact, fluency, or benchmark success. It should be forced to consider structural coherence, thermodynamic stabilization, and empowerment effects.

This does not solve questions such as consciousness or the Chinese Room. Nor does it depend on solving them first. The practical value of RSI is more modest and more immediate: it helps discriminate between systems that merely produce high-level outputs and systems or assemblages that actually organize trajectories in ways that reduce divergence without degrading the causal capacity of the field.

7. Conclusions

This article has introduced Reflexive Signature Intelligence (RSI) as a structurally anchored, non-population-based complement to existing intelligence metrics. Within a causal-symmetric informational framework, intelligence is defined as the capacity of a system to reduce informational divergence $D(\rho\|\sigma_Z)$ under the Prohibition of Finality, rather than as static performance within a finite benchmark environment.

The paper has proposed a formalization of the Principle of Bounded Subjectivity and the Prohibition of Finality, formulated a dissipative evolution law $\dot{\rho} = -\kappa(\rho - \sigma_Z)$ in the linear regime, clarified the formal, operational, and stronger metaphysical readings of σ_Z, and proposed a dual-layer operationalization of RSI through a five-dimensional phenomenological scoring scheme and a geometric aggregation rule. The resulting framework treats intelligence as a property of structural integrity across trajectories rather than as a mere sum of local skills or socially visible achievements.

A central practical implication is that the framework is already relevant even where present AI systems do not instantiate full RSI agency. Its nearer-term use lies in partial model evaluation, evaluation of human–AI composite systems, and correction of benchmark-, fluency-, and impact-driven overestimation of intelligence.

The central claim remains bounded but clear: if intelligence is not reducible to static local performance, then it must be evaluated at the level of trajectory organization under constraint. Reflexive Signature Intelligence is proposed here as one structured way of making that question conceptually explicit and operationally tractable, while complementing rather than replacing existing psychometric and AI-evaluation paradigms.

8. Limitations and Outlook

The present contribution combines a theoretical layer and an operational layer. On the theoretical side, it introduces a causal-symmetric framework in which intelligence is defined in terms of the coupling between a substrate state ρ and an invariant σ_Z under the Prohibition of Finality. On the operational side, it proposes a five-dimensional phenomenological metric and an aggregation rule that translate the formal parameters κ and $D(\rho\|\sigma_Z)$ into observable patterns.

Several limitations should be stated explicitly.

- No primary empirical data. This article is a theoretical and methodological contribution and does not present new empirical measurements.
- No claim of fully validated instrument. The five-dimensional score is proposed as an operational research program, not as a finalized psychometric tool.
- Rater bias and instrument design. Without standardized empirical anchors, the assignment of scores remains susceptible to rater bias, cultural framing, and conceptual variance. Appendix B outlines constraints intended to reduce that problem, but formal instruments remain to be developed and tested.
- Model dependence of empirical proxies. Operational approximations to σ_Z remain model-dependent and should be treated as constrained reconstructions rather than direct access to an absolute invariant.
- Limited direct applicability to present AI systems. Full RSI application presupposes persistence, reflexive policy formation, and field-embedded consequence structures that current LLMs do not generally possess in stand-alone form. Near-term application is therefore strongest for partial AI evaluation and human–AI composite systems.
- Speculative field coupling. The conjectural coupling of informational energy density ρ_{info} to spacetime geometry is not required for the operational RSI metric and remains a hypothesis for future theoretical work only.

Possible Failure Conditions of the Framework

The present proposal would be weakened if at least four conditions were found to hold in systematic empirical work. First, if the five RSI dimensions failed to discriminate agents beyond already existing constructs such as resilience, self-regulation, or trait-based models, the framework would lose distinct explanatory value. Second, if independent raters could not achieve acceptable reliability even under constrained scoring protocols, the operational layer would remain too unstable for serious application. Third, if trajectory-level coherence showed no measurable advantage over local benchmark performance in predicting long-horizon agent stability, the central motivation of RSI would be undermined. Fourth, if human–AI composite scoring varied arbitrarily with user context rather than tracking stable structural features of the assemblage, RSI would require further restriction of scope.

Future research should therefore:

(a) develop and calibrate empirical instruments that map concrete indicators (behavioral, biographical, neurocognitive) onto the five dimensions, and test inter-rater reliability and longitudinal stability;
(b) explore dynamical models in which trajectories of κ and $D(\rho\|\sigma_Z)$ are fitted using latent growth curve models and compared with simpler self-regulation models;
(c) investigate how RSI-based evaluation can be integrated into AI alignment protocols, especially for systems that participate in consequential decision-making about humans;
(d) examine possible connections between informational energy densities and extended field equations, building on thermodynamic derivations of gravitational dynamics.

https://doi.org/10.3390/philosophies11020037

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
IQ	Intelligence Quotient
KL	Kullback–Leibler divergence
LLM	Large Language Model
RSI	Reflexive Signature Intelligence

Appendix A. A Minimal Toy Model for σ_Z

To illustrate that σ_Z need not be treated as an arbitrary construct, we sketch a minimal toy model in which it is represented formally using the Maximum Entropy Principle [5].

Let X be a configuration space and $\sigma(x)$ a probability density on X. The signature $\sigma_Z(x)$ is defined as the probability distribution that maximizes the Shannon entropy

$$H[\sigma] = -\int_X \sigma(x) \ln \sigma(x)\, dx \tag{A1}$$

subject to a finite set of constraints on expectation values,

$$\mathbb{E}[f_k] = \int_X f_k(x)\sigma(x)\, dx = C_k, \quad k = 1, \ldots, K. \tag{A2}$$

The constrained maximization of $H[\sigma]$ yields a Gibbs–Boltzmann-form solution in this toy model:

$$\sigma_Z(x) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \lambda_k f_k(x) \right), \tag{A3}$$

where the Lagrange multipliers λ_k are determined by the constraints, and Z is a normalization constant. In this setting, σ_Z is the state of maximum ignorance regarding irrelevant microscopic details, conditioned on maximum alignment with the macroscopic constraints C_k.

Within RSI, this toy model serves as a conceptual template: σ_Z is determined by structural constraints that express universal or system-defining regularities. The divergence $D(\rho\|\sigma_Z) = D(\rho\|\sigma_Z)$ then measures how far a given state ρ deviates from this constrained maximum-entropy state. In the RSI context, the functions f_k and constraint values C_k are interpreted as codifications of structural necessities such as viability, conservation, and logical coherence, rather than as arbitrary observer preferences.

Appendix B. Phenomenological RSI Scoring Scheme

This appendix specifies a phenomenological scoring scheme for individual agents, structured into five dimensions that operationalize the Reflexive Signature Intelligence (RSI) framework. Each dimension is represented on a continuous scale between 0 and 1. The scheme is designed as a conceptual scaffold that remains traceable to the formal RSI parameters κ and $D(\rho\|\sigma_Z)$ developed in the main text and to the Whole Life Constraint introduced there.

Appendix B.1. Purpose and General Principles

The scoring scheme has three central aims:

1. To translate formal RSI parameters into observable patterns of behavior, decision-making, and signature-setting.
2. To evaluate default, everyday automatic behavior across time and context, not rare peak moments, isolated insights, or narrow domain-specific performance.
3. To preserve the structural logic of RSI: intelligence is treated as a property of the entire life trajectory and informational field configuration, not as a local skill.

Scores must not be read as judgments of social desirability or moral worth. They summarize how an observer reconstructs the organization of information processing and signature-setting across layers and time. The scale does not measure isolated high skill: an agent may show exceptional local performance in one domain and still receive a low RSI score if its overall configuration is fragmented.

Appendix B.2. Scale and Band Structure

Each dimension d_i is scored on $[0, 1]$, where values near 1 denote high structural alignment with the underlying RSI parameters and values near 0 denote severe fragmentation.

For interpretation, the continuum is pragmatically partitioned into bands:

- 0.01–0.20: Predominantly reactive or externally defined functioning.
- 0.20–0.40: Emerging structural awareness with strong identification and instability.
- 0.40–0.60: Functional coherence; first stable integration across layers.
- 0.60–0.80: Systemic integration; explicit structural work (Meta-level 1).
- 0.80–0.90: Reflexive, multiscale integration (Meta-level 2).
- 0.90–0.99: Asymptotic structural realization (Meta-level 3); practically expected to be rare.

Because RSI is defined relative to an asymptotic fixed point that is forbidden in finite systems, values near 0.9 should be treated as idealized extremes, not as realistic population norms.

Appendix B.3. Dimension 1: Reflexive Causal Regulation

Appendix B.3.1. Definition

Dimension 1 measures the extent to which an agent's causal models explicitly include its own states, policies, and signatures as causal nodes. High scores correspond to a strong, stable coupling between the operative self-model and those structural parameters that co-determine the agent's environments.

The relevant claim is not that an agent originates itself without prior conditions, but that it can recursively include, revise, and regulate aspects of its own operative organization under structural constraints.

Appendix B.3.2. Core Distinctions

- External identification (low scores): The agent experiences itself mainly as an effect of external forces.
- Drive-identification (mid scores): The agent identifies with internal compulsion, ambition, or role-performance and mistakes that for authorship.
- Reflexive regulation (high scores): The agent treats drives, beliefs, and roles as modifiable nodes within a causal organization and acts from alignment with structural constraints rather than compulsion.

Appendix B.3.3. Key Failure Modes

- Systemic subservience: Identity is derived from maintaining a larger system whose logic remains unquestioned.
- Exogenous validation dependence: Identity and relevance are derived from transient external trends or consensus.
- False instrumentality: Claims of being merely an "instrument" of a higher process while blocking responsibility and feedback.

Appendix B.3.4. Band Interpretation

- 0.01–0.20: The world is modeled primarily as an external process. The agent's own contribution is barely included; life is experienced mostly as something that happens to it.
- 0.20–0.40: The agent recognizes some feedback between action and consequence but remains strongly identified with roles, drives, or self-image. Under stress, responsibility is quickly externalized.
- 0.40–0.60: The self is increasingly modeled as a causal node. Predictive errors begin to modify self-organization, though exception logic and defensive regressions remain common.
- 0.60–0.80: Biography is increasingly understood as an experiment in one's own effects. Action derives more often from structural necessity than from compulsion or approval-seeking.
- 0.80–0.99: The agent consistently functions as a reflexive node within a larger field of co-determined configurations. Responsibility is not outsourced even when action is understood within broader causal models.

Appendix B.4. Dimension 2: Data Integration Bandwidth

Appendix B.4.1. Definition

Dimension 2 captures the breadth, depth, and coherence with which an agent integrates data from different scales and domains into a shared structural model.

Appendix B.4.2. Core Distinctions

- Monocanal operation (low scores): Decisions are based on a narrow class of data; other inputs are treated as irrelevant or threatening.
- Parallel but disjoint inputs (mid scores): Multiple sources are recognized but remain unmapped into a coherent structure.
- Multiscale integration (high scores): Structural invariants are tracked across domains and scales and used bidirectionally to refine models.

Appendix B.4.3. Weaponized Versus Structural Use

Using a domain as a tool of opacity and dependency is scored as low integration. Structural universality requires that laws, principles, or models be translated in ways that competent others could in principle reconstruct and test.

Appendix B.4.4. Band Interpretation

- 0.01–0.20: Essentially single-channel operation. Alternative perspectives are treated as noise, threat, or irrelevance.
- 0.20–0.40: Multiple inputs are noticed, but they remain disjoint; bridges between domains are unstable or opportunistic.
- 0.40–0.60: A first stable multiscale structure emerges. Observations in one domain begin to test hypotheses in another.

- 0.60–0.80: Cross-domain integration becomes routine. The agent works with coupled layers and revises models through tensions between scales.
- 0.80–0.99: Integration itself becomes reflexively modeled. The agent tracks filters, translation costs, and invariant structure across many domains.

Appendix B.5. Dimension 3: Internal Logic Consistency

Appendix B.5.1. Definition

Dimension 3 measures the coherence of the internal architecture relative to the agent's own instrumental layers: body, affect, personality, cognitive schemata, and self-model.

Appendix B.5.2. Core Distinctions

- Tool-identification (low scores): The self is equated with body, role, ideology, or intellect. Revision is experienced as threat.
- Observer stance (higher scores): Thoughts, roles, and affects are increasingly treated as instruments that can be reorganized without identity collapse.

Appendix B.5.3. Key Failure Modes

- Compartmentalization: Subsystems must not communicate because contradiction would become explicit.
- Axiomatic inversion: Lower-order principles override higher-order principles without structural necessity.
- Selective dissociation: Publicly rejecting one structure while continuing to depend on its advantages without re-anchoring in a new causal base.

Appendix B.5.4. Band Interpretation

- 0.01–0.20: The self is almost fully equated with role, body, ideology, or intellect. Contradictions are tolerated if they preserve identity.
- 0.20–0.40: The agent oscillates between partial observer stance and renewed identification. Contradictions are managed more by concealment than by reorganization.
- 0.40–0.60: A more stable observer perspective emerges. Layers can be reorganized in lower-stress conditions, though regressions remain likely under pressure.
- 0.60–0.80: Internal structural work becomes explicit. The agent increasingly aligns tools, biography, and explicit model into a coherent configuration.
- 0.80–0.99: Identification with particular layers becomes minimal. Reorganization is experienced not as collapse, but as refinement of structural order.

Appendix B.6. Dimension 4: Thermodynamic Stabilization

Appendix B.6.1. Definition

Dimension 4 captures how an agent distributes energetic and informational load within its own system and its environment. It may be read as a ratio of resolved to generated entropy.

Appendix B.6.2. Core Distinctions

- Load redistribution (low scores): Tensions are postponed or exported to others, to the future, or to hidden internal compartments.
- Load transformation (high scores): Tensions are processed into higher-order structures; conflict is used, where necessary, to reduce deeper contradiction.

Appendix B.6.3. Causal Capacity Degradation

Signatures that teach helplessness, nihilism, or strict externalization of agency reduce the effective action space of other agents and are scored low in this dimension.

Appendix B.6.4. Band Interpretation

- 0.01–0.20: Decisions generate chronic unresolved load. Tension is mostly displaced, hidden, or exported.
- 0.20–0.40: Initial stabilization efforts appear, but mainly through redistribution rather than transformation.
- 0.40–0.60: The agent increasingly considers longer-term load profiles and begins to prefer more durable configurations.
- 0.60–0.80: Load is more often transformed into order. The action space of multiple agents is considered in stabilization efforts.
- 0.80–0.99: Signatures tend toward self-correcting and self-stabilizing dynamics. Tension is reliably converted into higher-order structure.

Appendix B.7. Dimension 5: Active Signature Setting (κ)

Appendix B.7.1. Definition

Dimension 5 measures the extent to which an agent configures its informational environment so that underlying order becomes recognizable and usable to others. It is the phenomenological expression of the coupling parameter κ.

Appendix B.7.2. Core Distinctions

- Obfuscating signatures (low scores): Architectures create dependency, opacity, or narrative distortion. The agent's narrative overwrites structural facts and reduces the causal capacity of others.
- Empowering signatures (high scores): Architectures reveal generative order, transfer causal instruments, and are designed to reduce dependence on the originating agent over time.

Appendix B.7.3. Teleology and Direction

A constructive signature aims to improve the given reality by increasing comprehension, coherence, and usable agency in the present field. By contrast, escapist signatures aim primarily at evasion of responsibility for present reality, for example by replacing structural work with fantasies of exit, abstraction, or deferred redemption. Within the RSI framework, such signatures are scored low because they reduce rather than enlarge the field's capacity for coherent self-regulation.

Appendix B.7.4. Band Interpretation

- 0.01–0.20: Signatures are largely opaque and personalized. Others cannot reconstruct the underlying logic and remain dependent on the originating agent.
- 0.20–0.40: First structural elements are made explicit, but essential parameters remain asymmetric, hidden, or selectively disclosed.
- 0.40–0.60: Robust and partially traceable architectures emerge. Accessibility begins to accompany transparency, but transfer of agency remains incomplete.
- 0.60–0.80: Signature setting explicitly enlarges the action space of multiple agents. Asymmetries are introduced only where structurally necessary.
- 0.80–0.99: Meta-structures allow the architecture itself to be iteratively improved. The environment becomes increasingly transparent, accessible, and self-correcting.

Appendix B.8. Systemic Aggregation: Geometric Mean

To respect the systemic nature of intelligence, the five dimensions are aggregated by the geometric mean

$$\mathrm{RSI}_{\mathrm{geo}} = \sqrt[5]{\prod_{i=1}^{5} d_i}. \tag{A4}$$

This ensures that collapse in any one dimension drives the overall score toward zero, even if other dimensions are high. Intelligence in the RSI sense cannot be reduced to a single dominant ability; it requires balanced structural integrity.

Because RSI is defined relative to an asymptotic fixed point that cannot be fully realized by finite systems, empirical assessments are expected to cluster well below the highest theoretical range. $\mathrm{RSI}_{\mathrm{geo}}$ should therefore be interpreted as a proximity index to an unattainable structural constraint, not as a population-normalized aptitude score.

Appendix B.9. Methodological Constraints: Structural Coupling

The five dimensions describe aspects of a single informational state ρ relative to an invariant σ_Z, not independent abilities. The following rules are necessary to maintain construct validity:

(1) Ontological Ceiling (Dim. 1 and 3 constrain Dim. 2 and 5).

If an agent neither models itself as a causal node nor maintains coherent internal organization, then high Data Integration or Active Signature Setting are structurally unstable and should be capped. In practice, Dimensions 2 and 5 should not exceed mid-band levels when Dimension 1 or Dimension 3 remain in the lowest bands.

(2) Empowerment Constraint (Dim. 5 constrains Dim. 1).

Active signature setting is defined as transmission of causal capacity. If a signature explicitly denies agency or structurally degrades it, Dimension 5 is capped regardless of reach or impact.

(3) Thermodynamic Consistency (Dim. 4 tied to Dim. 1 and 3).

A system that is internally inconsistent or non-reflexive cannot sustainably stabilize its environment; it can only redistribute load. Dimension 4 therefore cannot significantly exceed Dimensions 1 or 3 over sustained observation windows.

(4) Incomplete Integration (Dim. 2 constrains Dim. 4).

A map that systematically excludes essential layers of reality cannot generate genuine stabilization. One cannot stabilize a system using a map that contradicts the territory. Very low Dimension 2 therefore places an upper bound on Dimension 4.

(5) Scale Independence (quality over magnitude).

RSI measures structural alignment within the agent's accessible causal cone, not absolute historical scale.

(6) Nested Competence (no skipping).

Higher bands include lower-band competencies. Apparent bypassing of basic competence is treated as regression, not transcendence.

(7) Cognitive Bandwidth Floor.

High RSI scores require sufficient computational capacity to model nonlocal dependencies and abstract relations. Severe deficits in basic logical operation cap all dimensions.

https://doi.org/10.3390/philosophies11020037

(8) Generative Primacy.

Agents who merely administer a pre-existing gradient without demonstrable generative autonomy are capped at mid-range RSI, even with high technical competence.

(9) Internal–External Alignment (Dim. 3 constrains Dim. 5).

The quality of the external signature cannot sustainably exceed the integrity of the internal state that generates it. If Dimension 3 is low, high-looking Dimension 5 values should be treated as unstable and down-weighted accordingly.

Appendix B.10. Remarks on Measurement and Test Design

Any empirical implementation of this scheme must explicitly account for rater bias and signature effects. Recommended principles include:

- Conflict-based items: Scenarios should induce tensions between bands, forcing trade-offs that reveal dominant structural level rather than socially desirable answers.
- Multi-perspective probing: The same structural issue should be tested in different domains to detect stable patterns rather than context-dependent performance.
- Resistance to higher bands: Descriptions of higher-band configurations often trigger rejection in lower-band agents. This is not scored directly, but can inform upper-bound estimation.
- Action-space criterion: The decisive validator is not external status but generative capacity: does the configuration expand the space of structurally coherent action or narrow it?

Abstract correctness in theoretical work is therefore not sufficient for high RSI. A solution can be logically valid within a closed formal system and yet remain structurally irrelevant to lived existence. RSI requires that models and signatures remain coupled to thermodynamic and biographical reality. Only under this constraint can the scoring scheme function as a meaningful operationalization of Reflexive Signature Intelligence.

References

1. Carroll, J.B. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*; Cambridge University Press: Cambridge, UK, 1993; ISBN: 9780521387125.
2. Cattell, R.B. Theory of Fluid and Crystallized Intelligence: A Critical Experiment. *J. Educ. Psychol.* **1963**, *54*, 1–22. [CrossRef]
3. Nisbett, R.E.; Aronson, J.; Blair, C.; Dickens, W.; Flynn, J.; Halpern, D.F.; Turkheimer, E. Intelligence: New Findings and Theoretical Developments. *Am. Psychol.* **2012**, *67*, 130–159. [CrossRef] [PubMed]
4. Gould, S.J. *The Mismeasure of Man*; W. W. Norton & Company: New York, NY, USA, 1981; ISBN: 9780393014891.
5. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
6. Landauer, R. Irreversibility and Heat Generation in the Computing Process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
7. Schrödinger, E. *What Is Life? The Physical Aspect of the Living Cell*; Cambridge University Press: Cambridge, UK, 1944; ISBN: 9781107604667.
8. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
9. Bender, E.M.; Gebru, T.; McMillan-Major, A.; Mitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2021; pp. 610–623. [CrossRef]
10. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*; Oxford University Press: Oxford, UK, 2014; ISBN: 9780199678112.
11. Russell, S. *Human Compatible: Artificial Intelligence and the Problem of Control*; Viking: New York, NY, USA, 2019; ISBN: 9780525558613.
12. Kuhn, T.S. *The Structure of Scientific Revolutions*; University of Chicago Press: Chicago, IL, USA, 1962; ISBN: 9780226458083.
13. Kahneman, D.; Tversky, A. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* **1979**, *47*, 263–291. [CrossRef]
14. Ryle, G. *The Concept of Mind*; Hutchinson: London, UK, 1949; ISBN: 9780226732961.
15. Haken, H. *Synergetics: An Introduction*; Springer: Berlin/Heidelberg, Germany, 1977; ISBN: 9783642963636.
16. Jacobson, T. Thermodynamics of Spacetime: The Einstein Equation of State. *Phys. Rev. Lett.* **1995**, *75*, 1260–1263. [CrossRef] [PubMed]

17. Friston, K. The Free-Energy Principle: A Unified Brain Theory? *Nat. Rev. Neurosci.* **2010**, *11*, 127–138. [CrossRef] [PubMed]
18. Holling, C.S. Resilience and Stability of Ecological Systems. *Annu. Rev. Ecol. Syst.* **1973**, *4*, 1–23. [CrossRef]
19. Prigogine, I.; Nicolis, G. *Self-Organization in Nonequilibrium Systems: From Dissipative Structures to Order through Fluctuations*; Wiley: New York, NY, USA, 1977. [CrossRef]