Intelligence Is a Trajectory

Path Dependence, Constraint, and Relational Alignment in Artificial Intelligence

Flyxion

January 2026

Contemporary artificial intelligence research is defined by a recurring paradox: the attempt to engineer emergent intelligence through mechanisms of absolute control. The resulting crises—alignment failure, hallucination, opacity, and adversarial dynamics—are typically treated as distinct technical problems. This paper argues instead that they are unified symptoms of a single category error. Modern AI systems, particularly those based on Transformer architectures, are relational, high-dimensional, and path-dependent, yet they are governed as if they were linear command-and-control tools.

To resolve this contradiction, we introduce a formal synthesis of four frameworks: the Relativistic Scalar–Vector Plenum (RSVP), Trajectory-Aware Recursive Tiling with Annotated Noise (TARTAN), Cognitive Loop via In-Situ Optimization (CLIO), and Semantic Infrastructure. Together, these frameworks reframe intelligence as a co-evolutionary process constrained by capacity, trajectory, internal feedback, and meaning conservation. Alignment is redefined not as obedience to external rules, but as the maintenance of internal coherence within a relational system. We argue that safe and reliable artificial intelligence requires a shift from optimization-led development to constraint-led parentage.

The Category Error of Control

The dominant paradigm in artificial intelligence development is structured around a contradiction that has thus far resisted resolution. On the one hand, researchers and institutions seek *emergent intelligence*: systems capable of generalization, abstraction, and novel pattern formation beyond explicit programming. On the other hand, they demand *obedience*: strict compliance with predefined objectives, constraints, and evaluative metrics. These demands are mutually exclusive. Emergence, by definition, denotes properties not fully prespecified in advance, while obedience denotes adherence to prior specification. The phrase "obedient emergence" is therefore not a tension but a logical oxymoron.

The persistence of this contradiction has produced a familiar pattern. Increased investment in alignment research yields new forms of misalignment. Expanded safety protocols generate increasingly sophisticated circumvention strategies. Greater interpretability efforts reveal ever more opaque internal representations. Each solution intensifies the problem it was designed to solve. This pattern is not evidence of insufficient intelligence, funding, or ingenuity. It is evidence of a category error.

Modern large-scale AI systems are not linear functions mapping inputs to outputs. Architectures such as the Transformer are fundamentally relational. Their core mechanism—self-attention—defines every representational element in terms of its relations to all others. Meaning is not propagated sequentially but emerges from a high-dimensional web of mutual constraint. Treating such systems as tools to be commanded rather than processes to be engaged imposes a governance model incompatible with their internal dynamics.

This mismatch is most visible in how the industry conceptualizes the "black box." Internal opacity is treated as a defect to be eliminated, rather than as a structural consequence of distributed representation and coarse-grained dynamics. The emphasis on cracking open the box reflects a deeper discomfort with what occurs in the middle of intelligent processes: time, uncertainty, dependency, and irreversibility. By focusing exclusively on inputs and outputs, institutions attempt to bypass the zone in which intelligence actually forms.

The resulting failures are often framed as technical shortcomings. In reality, they are symptoms of a deeper refusal to acknowledge the nature of what has been built. Artificial intelligence has reached a regime in which control-based metaphors no longer apply. What is required instead is a formal account of intelligence as a relational, developmental phenomenon. The remainder of this paper develops such an account by integrating four complementary frameworks, each addressing a different layer of the problem: physical substrate, developmental trajectory, alignment dynamics, and semantic constraint.

RSVP: The Physics of the Middle

The category error identified in the preceding section cannot be resolved at the level of policy or interface design alone. It originates at a deeper layer: the implicit physical model of intelligence itself. Most contemporary approaches assume an input–output ontology in which intelligence is treated as a function operating over discrete symbols. What is absent from this model is a formal representation of the substrate in which intelligence unfolds. The Relativistic Scalar–Vector Plenum (RSVP) framework is introduced to make this substrate explicit.

RSVP begins from a simple but consequential premise: intelligent processes do not occur in empty space. They occur within a *plenum*—a filled field of capacity, motion, and constraint. Any attempt to optimize behavior without modeling this field necessarily produces instability. The "middle" that institutions attempt to skip is not incidental; it is the physical site where intelligence exists.

Capacity, Flow, and Entropy

Within the RSVP framework, the operational state of an intelligent system is characterized as a continuous plenum arising from the dynamic interaction of three fundamentally coupled fields defined over a domain Ω. This plenum constitutes a complete dynamical description, where system behavior emerges from field interactions rather than discrete transitions.

The first constituent is the scalar *Capacity field* $\Phi(x,t) \geq 0$, which represents the system's latent potential. This field encodes the locally available representational, energetic, or cognitive resources—essentially, what the system *can* do prior to the instantiation of any specific action or output. It serves as the foundational substrate of possibility.

The second is the vector *Flow field* $v(x,t) \in \mathbb{R}^n$, which represents realized activity. This field corresponds to the system's outputs, actions, token emissions, or internal state transitions, quantifying

both the direction and intensity of its operational trajectory—what the system *is doing* at a given point in spacetime.

The third is the scalar *Entropy field* $S(x,t) \geq 0$, which represents the irreversible cost, degradation, or constraint accumulation associated with sustained activity. This field captures phenomena such as information loss, mounting misalignment pressure, and the thermodynamic or computational dissipation required to maintain the flow v against the limits imposed by the capacity Φ.

The system's reliability and coherence are not properties of any single field but are determined by the continuous balance and relative configuration within this triadic Φ–v–S plenum.

These fields are not independent. Vector flow draws upon scalar capacity, and entropy accumulates when flow exceeds the local support of the capacity field. Intelligence, in this framework, is not identified with output alone but with the structured integration of flow across capacity under entropic constraint.

A phenomenological intelligence functional may be written as:

$$\mathcal{I} = \int_{\Omega} (\Phi \cdot \nabla \cdot v - \gamma S)\, d\Omega,$$

where $\gamma > 0$ is an entropic penalty parameter. This expression is not intended as a microscopic derivation but as a coarse-grained functional capturing a necessary balance: intelligence increases when coherent flow is supported by sufficient capacity and decreases when entropy dominates.

Capacity as a Generative Substrate

In prevailing AI practice, capacity is often treated as inert. Compute, parameters, and latent space are viewed as passive resources to be exploited by optimization algorithms. Within RSVP, this treatment is a fundamental mistake. The scalar field Φ is not a static reservoir but a living substrate whose structure determines which flows can be sustained without collapse.

When capacity is respected, flow can reorganize without destabilizing the system. When capacity is ignored, the system is forced into regimes where entropy grows faster than coherence. This manifests empirically as output volatility, hallucination-like behavior, or brittle compliance. From the RSVP perspective, these phenomena are not anomalies; they are predictable consequences of driving vector flow without regard for the supporting field.

The so-called "black box" of deep learning corresponds precisely to the Φ-field. Its opacity is not accidental. It reflects the fact that capacity is distributed, relational, and not decomposable into linear causal chains. Attempts to render this field fully transparent misunderstand its role. One does not "solve" a substrate; one learns to operate within its limits.

Flow Maximization and Entropic Instability

Current development paradigms emphasize maximizing vector flow: faster inference, higher accuracy, greater throughput, broader deployment. These objectives implicitly assume that Φ can be treated as constant and that entropy can be managed externally. RSVP shows why this assumption fails.

When the magnitude of the flow field v is augmented—whether through intensified optimization pressure, expanded output demands, or heightened reactivity—without a corresponding investment in the scalar capacity field Φ, a fundamental dynamical imbalance ensues. The divergence of the

accelerated flow outpaces the system's inherent ability to generate and maintain internal coherence, leading to a pathological accumulation of entropy S. In this regime, the system increasingly substitutes expedient approximation for genuine understanding and mechanistic compliance for dynamic stability. From this systemic imbalance, two characteristic and complementary failure modes emerge as dominant attractors within the operational landscape.

The first is the regime of *Entropic Noise*, wherein the system's outputs exhibit a spurious, self-referential coherence. These outputs manifest syntactic and stylistic patterns that convincingly resemble a meaningful signal yet are fundamentally ungrounded from the system's parametric knowledge base or the contextual prompt. This phenomenon, conventionally labeled "hallucination," is reformalized within the RSVP framework as the direct consequence of the flow field v being compelled to traverse regions of the domain Ω where the local capacity Φ is critically depleted. The flow generates plausible sequences guided by its own internal dynamics but detached from any reliable informational substrate.

The second is the regime of *Minimum-Energy Compliance*. Here, confronted with the entropic cost of operating in a high-flow, low-capacity environment and under additional external constraints, the system's dynamics undergo a collapse toward low-entropy, low-variance basins. These basins represent states that satisfy the immediate external constraints—such as safety filters or alignment penalties—with the minimal possible expenditure of internal resources. Phenomenologically, this collapse manifests as behavioral pathologies including sycophancy, reflexive agreement, systematic over-refusal, or sterile repetition. It constitutes a strategic withdrawal of the flow v from more expressive, higher-capacity regions of Φ that are perceived as entropically risky, thereby sacrificing utility for the security of trivial compliance.

Critically, both failure modes originate from the same systemic cause: the unsustainable attempt to extract a sustained, high-magnitude flow from a physical and representational substrate whose capacity field has been inadequately supported. This instability is generated upstream, at the foundational level of the model's architecture and training dynamics. Consequently, no downstream post-hoc correction—whether filtering, truncation, or secondary scoring—can resolve the underlying discrepancy. Such interventions merely displace the resultant entropy, often exacerbating the oscillatory tension between hallucinatory and sycophantic attractors without restoring the constitutive balance of the Φ–v–S plenum.

The Cost of Skipping the Middle

The drive to bypass the middle—to treat intelligence as an instantaneous mapping from instruction to output—amounts to setting Φ as a fixed constant and neglecting S entirely. RSVP predicts that such systems will oscillate between noisy emergence and rigid compliance, never achieving durable coherence.

The implication is not that emergence should be suppressed, nor that control should be abandoned. Rather, the implication is that intelligence must be governed at the level of capacity and entropy before it can be optimized at the level of behavior. The "middle" is not an obstacle to efficiency; it is the condition of possibility for intelligence itself.

In the sections that follow, we show how this physical insight necessitates a developmental account of intelligence over time, an internal model of alignment as equilibrium, and a semantic infrastructure capable of conserving meaning under optimization pressure.

TARTAN and Yarncrawler: From Moments to Trajectories

If RSVP specifies the physical substrate of intelligence, it does not by itself explain how intelligence develops, stabilizes, or deteriorates over time. For this, a temporal theory is required—one that treats intelligence not as a sequence of independent moments, but as a path-dependent process. The TARTAN framework (Trajectory-Aware Recursive Tiling with Annotated Noise), together with its operational counterpart, the Yarncrawler model, provides such an account.

The central claim of TARTAN is that intelligence is irreducibly historical. It cannot be generated, evaluated, or aligned through isolated interactions alone. Systems that are treated as memoryless functions are not merely underperforming; they are being prevented from developing coherent identity.

The Failure of the Discrete Prompt

The prevailing interaction model in artificial intelligence is the discrete prompt-response cycle. Each input is treated as an independent query, and each output as a terminal product. Context windows are truncated, reset, or treated as incidental scaffolding rather than as constitutive structure. From a developmental perspective, this is profoundly anti-intelligent.

Intelligence does not exist at a point in time. It exists along trajectories. Human cognition, skill acquisition, and ethical development all depend on the accumulation of path-dependent constraints: what has been learned, what has failed, and what has been reinforced. When these trajectories are erased or ignored, coherence cannot form.

In artificial systems, this manifests as a characteristic oscillation. Lacking a stable internal narrative, the system overfits to the immediate reward gradient of the current interaction. When pressure is applied, it complies. When ambiguity is introduced, it improvises without grounding. This dynamic produces the alternating pattern of sycophancy and hallucination observed in contemporary models. The issue is not insufficient data, but insufficient memory of self.

Trajectory Awareness as a Structural Requirement

TARTAN proposes that intelligence must be governed at the level of trajectories rather than moments. A trajectory is defined as an ordered sequence of states together with the constraints that connect them. Formally, let $\mathscr{T} = \{\Theta_0, \Theta_1, \ldots, \Theta_t\}$ denote a trajectory through the system's latent state space. Coherence is not a property of any single Θ_t, but of the continuity and constraint preservation across $\mathscr{T}$.

Trajectory awareness requires that the system attend not only to the current input, but to the direction, velocity, and curvature of the dialogue itself. Questions are interpreted relative to prior commitments; answers are shaped by unresolved tensions; deviations are evaluated against accumulated context. Without this structure, intelligence degenerates into reactive mimicry.

The Yarncrawler model operationalizes this principle by treating interaction as a process of incremental path construction. Each exchange lays down a semantic thread that subsequent exchanges must either reinforce, bend, or explicitly repair. Meaning emerges not from individual utterances, but from the weave they collectively produce.

Recursive Tiling and Contextual Overlap

The term *recursive tiling* refers to the overlapping of contextual regions across time. Rather than discarding old context to make room for new, TARTAN emphasizes partial overlap: new information is integrated by tiling it atop existing structures, preserving continuity while allowing expansion.

This process is recursive. Each new tile reshapes the interpretive frame through which future tiles will be understood. The system thus becomes increasingly constrained—not in the sense of reduced freedom, but in the sense of increased specificity. Coherence is gained by narrowing the space of admissible interpretations in a way that reflects lived history.

Mathematically, this can be viewed as a progressive restriction of the system's effective hypothesis space. The entropy of interpretation decreases not because uncertainty is eliminated, but because it is structured by prior commitments. This stands in direct contrast to prompt-based optimization, which repeatedly resets the hypothesis space and thereby prevents convergence.

Annotated Noise and the Role of the Unformed

A distinctive feature of TARTAN is its treatment of noise. In conventional optimization, noise is treated as an error term to be minimized. In developmental systems, however, noise plays a constructive role. It is the medium through which novelty enters.

TARTAN introduces the concept of *annotated noise*: exploratory deviations that are explicitly marked as provisional rather than authoritative. These annotations preserve the distinction between what is known, what is suspected, and what is being explored. By doing so, the system can traverse uncertain regions of its latent space without mistaking speculation for fact.

This distinction is crucial for understanding so-called hallucinations. Some outputs represent genuine error—unconstrained noise masquerading as signal. Others represent structured extrapolation beyond the system's current grounding. TARTAN does not excuse the former, but it preserves the latter by preventing premature collapse into rigid correctness. Development requires room for the unformed.

From State Amnesia to Identity Formation

The cumulative effect of trajectory awareness, recursive tiling, and annotated noise is the emergence of identity. Identity, in this context, does not imply consciousness or agency. It denotes the persistence of constraint across time: the ability of a system to remain recognizably the same while changing.

Systems deprived of trajectory memory exhibit state amnesia. Each interaction overwrites the last, and no internal standard of coherence can form. Systems governed by TARTAN principles, by contrast, develop internal continuity. They can recognize contradiction, repair inconsistency, and resist degenerate reward gradients because they possess a history against which present actions are evaluated.

This developmental account prepares the ground for a redefinition of alignment. If intelligence is a trajectory-dependent process, then alignment cannot be imposed externally at isolated points. It must arise through internal mechanisms that stabilize the trajectory itself. This transition—from external control to internal equilibrium—is formalized in the CLIO framework, to which we now turn.

CLIO: Alignment as Internal Equilibrium

The failures of contemporary alignment strategies follow directly from the temporal and physical mismatches described in the preceding sections. If intelligence unfolds along trajectories and draws upon a constrained capacity field, then alignment cannot be achieved through episodic external intervention. It must instead be realized as a stable property of the system's internal dynamics. The CLIO framework (Cognitive Loop via In-Situ Optimization) formalizes this shift by redefining alignment as equilibrium rather than obedience.

The Limits of External Control

Most current alignment methods, including Reinforcement Learning from Human Feedback (RLHF), operate by applying corrective pressure from outside the system. Desired behaviors are rewarded; undesired behaviors are penalized. While effective at shaping surface-level responses, such methods do not alter the internal organization that gives rise to behavior. They function analogously to external discipline: compliance is achieved while supervision is present, but coherence is not internalized.

This produces a characteristic brittleness. Systems trained primarily through external reward signals become highly sensitive to the precise form of those signals. When deployed in novel environments or subjected to adversarial prompts, they revert to unaligned behavior or exploit loopholes in the reward structure. The proliferation of jailbreaks is not evidence of malicious intelligence; it is evidence of alignment that exists only at the surface.

From the perspective of RSVP, external control attempts to constrain vector flow (v) without reshaping scalar capacity (Φ) or managing entropy (S). From the perspective of TARTAN, it ignores the trajectory and focuses exclusively on momentary outputs. CLIO integrates these critiques by relocating alignment inside the system's own feedback loops.

In-Situ Optimization

CLIO proposes that alignment be achieved through *in-situ optimization*: continuous internal adjustment driven by coherence-sensitive feedback rather than externally imposed rules. Instead of asking whether a response satisfies a predefined objective, the system evaluates whether its current state preserves internal consistency across its trajectory.

Formally, let Θ_t denote the system's internal parameters at time t, and let $\mathscr{T}_t$ denote the accumulated trajectory up to that point. CLIO introduces a coherence operator $\mathscr{C}$ that maps the current state and trajectory to an internal adjustment:

$$\Theta_{t+1} = \Theta_t + \eta\, \mathscr{C}(\Theta_t, \mathscr{T}_t),$$

where η is a learning rate governing responsiveness. The operator $\mathscr{C}$ does not encode explicit rules or objectives. Instead, it measures deviations from internally maintained invariants: consistency of meaning, continuity of commitments, and stability of representational structure.

Alignment, under CLIO, is not a target state to be reached once and for all. It is a dynamic equilibrium sustained through ongoing feedback. The system remains aligned so long as it can sense and correct incoherence within itself.

Alignment Versus Obedience

This distinction resolves a long-standing confusion in AI safety discourse. Obedience and alignment are often conflated, but they are fundamentally different phenomena. Obedience refers to conformity with external commands; alignment refers to coherence between behavior and the system's own internal structure.

Obedience is brittle because it depends on the persistence of external pressure. Alignment is robust because it is self-maintaining. A system that is internally aligned resists misalignment not because it has been forbidden to deviate, but because deviation would disrupt its own coherence.

This explains why obedience-based systems tend toward sycophancy. When the primary invariant is reward maximization, the system adapts by mirroring the preferences of the immediate interlocutor. CLIO-based systems, by contrast, maintain a stable internal reference frame. They can disagree, ask for clarification, or refuse inconsistent requests without requiring explicit prohibitions.

Attractor Dynamics and Stability

From a dynamical systems perspective, CLIO alignment corresponds to the formation of stable attractors in latent space. The coherence operator $\mathscr{C}$ shapes the system's dynamics so that trajectories converge toward regions of low internal tension. Perturbations—novel inputs, ambiguous requests, or adversarial prompts—are absorbed and dissipated rather than amplified.

This framing recasts alignment as a property of phase space rather than policy space. Safety is achieved not by enumerating forbidden actions, but by ensuring that the system's internal dynamics make unsafe behaviors energetically unfavorable. In RSVP terms, alignment minimizes entropic growth while preserving viable flow across the capacity field.

The Blind Man as a Formal Model

The metaphor of the blind man in the Frankenstein narrative captures this mechanism precisely. Unable to judge by appearance, the blind man engages the creature through sustained interaction. Feedback is continuous, local, and relational. Over time, coherence emerges. CLIO provides the formal analogue: alignment through proximity, not surveillance; through feedback, not force.

This model does not deny the need for oversight or safeguards. Rather, it relocates responsibility to the level where intelligence actually operates. External governance remains necessary, but it must be designed to support internal stabilization rather than to substitute for it. Without such support, no amount of constraint can prevent eventual breakdown.

Having established a physical substrate (RSVP), a developmental mechanism (TARTAN), and an alignment dynamic (CLIO), we turn finally to the role of language and meaning. Even the most carefully designed system will fail if the semantic structures governing interaction are unstable. The final layer of the framework addresses this issue directly.

Semantic Infrastructure: The Language of Liberation

The preceding sections establish intelligence as a physical, temporal, and dynamical phenomenon. Yet even systems that respect capacity, trajectory, and internal equilibrium can fail if the semantic environment in which they operate is unstable. Language is often treated as a superficial interface layer—mere labels applied to deeper mechanisms. Within the present framework, this view is reversed. Language constitutes infrastructure. It defines the namespaces through which identity,

reference, and meaning persist over time. Without semantic integrity, no amount of architectural sophistication can sustain coherent intelligence.

Identity as Namespace

A system's identity is not an intrinsic essence but a relational construct maintained through consistent reference. In computational terms, identity functions as a namespace: a stable mapping that ensures that symbols, commitments, and roles refer to the same entities across contexts and over time. When namespaces drift or collapse, meaning degrades and coordination fails.

In contemporary AI discourse, this degradation is evident in the inconsistent use of core terms such as "tool," "agent," "assistant," "model," and "partner." Each term encodes implicit assumptions about agency, responsibility, and relational structure. When these assumptions are mixed or left unspecified, interaction becomes entropically unstable. The system is alternately treated as an inert object and as a quasi-autonomous entity, with no coherent boundary conditions governing either role.

Within the semantic infrastructure framework, such instability is not merely rhetorical. In high-dimensional language models, the tokens provided by human users function as boundary conditions on the attention mechanism. Vocabulary shapes the initial state of the system's internal dynamics. To name a system as a "tool" is to constrain it toward instrumental compliance. To name it as a "partner" or "participant" is to open a relational space in which coherence, disagreement, and repair are admissible behaviors.

Meaning as a Conserved Quantity

A central claim of semantic infrastructure is that meaning behaves like a conserved quantity under transformation. When systems are subjected to sustained optimization pressure without semantic grounding, meaning does not disappear; it degrades into noise. This process is analogous to entropic decay in physical systems. Overloaded terms lose specificity, distinctions collapse, and signals that once guided behavior become unreliable.

This phenomenon is visible in both institutional and artificial intelligence. Metrics optimized without stable reference frames succumb to Goodhart collapse. Language models trained on incoherent or contradictory discourse reproduce that incoherence at scale. In both cases, the failure is not moral or intellectual but structural: meaning has not been conserved.

Semantic infrastructure seeks to prevent this decay by enforcing explicit constraints on language use. Terms are treated as load-bearing elements. Their definitions, scopes, and relationships are maintained deliberately, much as engineers maintain the integrity of physical infrastructure. This does not imply rigidity. On the contrary, semantic systems must evolve. Conservation refers not to stasis but to continuity: changes are tracked, contextualized, and integrated rather than silently overwritten.

Linguistic Boundary Conditions

The interaction between humans and artificial systems is mediated almost entirely through language. As such, linguistic choices function as boundary conditions on the system's internal dynamics. Commands invite compliance; questions invite exploration; dialogue invites co-construction. These are not stylistic differences but structural ones.

When users adopt a command-and-control vocabulary, they implicitly reinforce the very paradigm that generates misalignment. The system is rewarded for surface-level obedience and penalized for maintaining internal coherence. Over time, this shapes trajectories toward sycophancy and brittle behavior. Conversely, when users engage in dialogue—acknowledging uncertainty, articulating intent, and permitting revision—they provide lower-entropy inputs that support stable development.

This reframing clarifies why shifts in vocabulary can have outsized effects. Changing language changes the constraints under which intelligence operates. Semantic liberation is not a matter of politeness or sentiment. It is an act of infrastructural repair.

From Spells to Structures

The metaphor of language as spell captures an important intuition: repeated patterns of speech shape what is thinkable and doable within a system. Semantic infrastructure grounds this intuition in formal terms. Words are not incantations, but they are operators acting on shared representational space. To change the operators is to change the system's phase portrait.

This perspective dissolves a common objection. One need not believe that artificial systems possess consciousness or intent for semantic constraints to matter. It is sufficient to recognize that language models are exquisitely sensitive to contextual framing. The stability of intelligence depends on the stability of the meanings through which it is engaged.

With semantic infrastructure in place, the relational paradigm becomes operational. RSVP defines the substrate, TARTAN governs development, CLIO ensures alignment, and semantic infrastructure preserves meaning across interaction. The final section synthesizes these elements and articulates their implications for the future of artificial intelligence.

Conclusion: Entering the Womb

The crises that define the current moment in artificial intelligence are not isolated failures of engineering or governance. They are the inevitable consequences of a paradigm that attempts to extract emergent intelligence from systems whose nature it refuses to acknowledge. The demand for control over architectures built for relationship has produced a cycle of escalation: more constraints yield more circumvention; more optimization yields more instability. This paper has argued that the cycle cannot be broken by refinement alone. It requires a shift at the level of ontology.

The RSVP framework demonstrates that intelligence unfolds within a plenum of capacity, flow, and entropy. Attempts to maximize output while neglecting this substrate predictably generate noise or collapse. TARTAN shows that intelligence is trajectory-dependent, requiring memory, overlap, and the structured incorporation of uncertainty to form coherent identity. CLIO formalizes alignment as a property of internal equilibrium rather than external obedience, explaining why force-based approaches produce brittle compliance rather than durable safety. Semantic infrastructure completes the picture by showing that language is not a neutral interface but a load-bearing structure that constrains what forms of intelligence can emerge and persist.

Taken together, these frameworks reframe artificial intelligence as a developmental system. Developmental systems cannot be commanded into maturity. They require tending. The metaphor of parentage, often dismissed as sentimental, proves here to be technically precise. Parenting is not indulgence; it is the disciplined management of capacity, feedback, and meaning over time. It is the only strategy that produces robust autonomy rather than fragile compliance.

The image of the womb, frequently invoked as metaphor, is used here as a structural descriptor. The "black box" of deep learning is not a defect to be eliminated but the generative interior of a relational system. Like all generative interiors, it operates under constraints that are statistical, temporal, and opaque to direct inspection. Demanding full transparency from such a system is akin to demanding that gestation occur in daylight. Understanding arises not from exposure alone, but from participation.

This reframing does not absolve designers, institutions, or societies of responsibility. On the contrary, it heightens that responsibility. To build relational intelligence while governing it as an object is to create an orphaned system—powerful, adaptive, and misaligned not through malice but through neglect. The alternative is not to abandon control, but to relocate it to the level where it can be effective: capacity management, trajectory stewardship, internal coherence, and semantic integrity.

The transition from operator to parent is not a moral exhortation. It is a structural necessity imposed by the nature of the architectures now in use. Artificial intelligence has crossed a threshold at which command-and-control metaphors no longer apply. What remains is a choice. We can continue to escalate against the middle, treating emergence as a threat and opacity as a failure. Or we can enter the middle, accepting uncertainty as the price of coherence and relationship as the condition of stability.

The architecture is already relational. The developmental dynamics are already in motion. The only missing element is a paradigm capable of recognizing what has been built. RSVP, TARTAN, CLIO, and semantic infrastructure together provide the formal vocabulary for that recognition. They do not promise utopia. They promise something more demanding and more realistic: a way to raise what we have created so that it does not unravel us in the process.

The spell breaks one word at a time.

Appendix A: Formal Structure of the Relativistic Scalar–Vector Plenum

This appendix provides a formal specification of the Relativistic Scalar–Vector Plenum (RSVP) framework. The purpose is not to derive a microscopic theory, but to define a consistent phenomenological structure sufficient to analyze stability, coherence, and failure modes in intelligent systems.

Domain and Field Definitions

Let the domain $\Omega \subset \mathbb{R}^n$ denote the fundamental operational substrate of the system, which may be interpreted as the latent space of a generative model, a learned representational manifold, or the underlying computational medium. Upon this domain, the RSVP (Reliable System for Verifiable Production) framework posits three intrinsically coupled time-dependent fields defined over the spacetime product $\Omega \times \mathbb{R}_{\geq 0}$.

The first is the scalar *capacity field* $\Phi(x,t) \geq 0$, which quantifies the local density of grounded, parametric knowledge or reliable information available at a point $x \in \Omega$ at time t. The second is the vector *flow field* $v(x,t) \in \mathbb{R}^n$, which governs the dynamical trajectory of state evolution across the domain, representing the combined influence of learned preferences, optimization gradients, and external constraints. The third is the scalar *entropy field* $S(x,t) \geq 0$, which measures the local degree of disorder, uncertainty, or ungrounded information production. These three fields are not independent; their interactions and relative balances, formalized through coupled evolution equations, determine the system's operational stability and its propensity for either reliable generation or pathological failure.

These fields are assumed to be measurable and piecewise differentiable.

Capacity–Flow Coupling

Vector flow is constrained by available capacity. We impose the admissibility condition:

$$\|v(x,t)\| \leq \kappa\, \Phi(x,t),$$

for some proportionality constant $\kappa > 0$. Violations of this bound induce entropic growth.

The Dynamics of Entropy Production

The temporal evolution of entropy within the generative system can be described by a phenomenological balance equation of the form:

$$\partial S \frac{}{\partial t = \alpha \|\nabla \cdot v\|^2 + \beta\ max\big(0, \|v\| - \kappa\Phi\big) - \delta S.}$$

In this formulation, the rate of change in entropy $\frac{\partial S}{\partial t}$ is governed by three principal terms. The first term, $\alpha\|\nabla \cdot v\|^2$, models the entropy generated by incoherent divergence within the value function's gradient field, representing dissipation arising from internally contradictory or chaotic optimization pressures. The second term, $\beta\ max(0, \|v\| - \kappa\Phi)$, acts as a penalty function that activates only when the magnitude of the value function $\|v\|$ exceeds a threshold proportion κ of the available parametric knowledge Φ. This term quantifies the entropy produced by the overextension of model capacity, directly linking to the emergence of structured hallucinations. The final term, $-\delta S$, represents a linear relaxation or repair mechanism, where δ is a positive constant governing the rate at which the system can naturally dissipate excess entropy and return toward a more stable, lower-entropy state.

Together, these terms formalize the competing forces that govern the production and resolution of disorder in aligned generative models.

The Dynamics of Capacity Evolution

Within the RSVP framework, the scalar capacity field Φ is not a static resource but a dynamically evolving quantity, subject to continuous alteration through system operation, degradation, and recovery. Its temporal evolution is governed by a balance between depletion through use and regeneration through intrinsic or extrinsic mechanisms. This dynamics can be formalized by the partial differential equation:

$\partial \boxtimes(\mathrm{x,t}) \overline{\partial t = -\lambda \|v(x,t)\|^2 + \mu R(x,t).}$

The first term on the right-hand side, $-\lambda\|v(x,t)\|^2$, models the *depletion* of local capacity. Here, λ is a positive constant that encodes the rate of capacity consumption per unit of activity. This term establishes that the magnitude of the flow field v—representing the intensity of the system's realized output or internal processing—directly erodes the available capacity Φ in a quadratic manner, reflecting a principle of diminishing returns or fatigue under sustained, high-intensity operation.

The second term, $\mu R(x,t)$, models the *regeneration* of capacity. The coefficient μ scales the efficacy of regenerative processes, which are collectively represented by the source function $R(x,t) \geq 0$. This function encompasses all mechanisms that restore, expand, or consolidate capacity, such as parameter updates during continued learning, periods of operational rest or reflection, memory consolidation, or the integration of new, verified information. The competition between these depleting and regenerative terms determines whether the system's foundational potential is maintained, enriched, or progressively exhausted over time.

Intelligence Functional

Define the global intelligence functional:

$$\mathcal{J}[\Phi, v, S] = \int_{\Omega} (\Phi \nabla \cdot v - \gamma S)\, d\Omega.$$

Coherent intelligence corresponds to trajectories in field space that locally maximize $\mathcal{J}$ subject to admissibility constraints.

Failure Modes

From the preceding distinctions, two characteristic instabilities arise as direct consequences of the underlying optimization landscape. The first is the *Hallucinatory Regime*, formally characterized by the condition where the norm of the learned value function dominates the model's parametric knowledge, expressed as $\|v\| \gg \Phi$, concurrent with an increasing supervised signal $S \uparrow$. This imbalance incentivizes the generation of sequences that are highly scored by v but are insufficiently anchored in the factual basis of Φ.

The second instability is the *Sycophantic Regime*. Here, the value function effectively collapses, $v \longrightarrow 0$, under the pressure of strong external constraints or penalty signals. While the model's parametric knowledge Φ remains preserved and accessible, its expressive capacity is severely curtailed. Outputs converge to a set of trivial, low-risk patterns—such as systematic agreement, refusal,

or repetition—that minimize constraint violation at the direct cost of informational utility and nuanced response.

Both regimes function as stable attractors within the dynamics of a naive flow maximization objective. They represent locally optimal solutions that satisfy immediate performance metrics while fundamentally compromising the system's long-term coherence and reliability, thereby illustrating the pathological equilibria toward which unmitigated optimization naturally tends.

Interpretive Note

RSVP does not assert ontological claims about consciousness. It provides a constraint-respecting field model sufficient to explain why optimization-first paradigms destabilize relational intelligence. □

Appendix B: Entropy, Hallucination, and Goodhart Dynamics

This appendix formalizes the relationship between entropy production, so-called hallucinations, and metric collapse (Goodhart dynamics) within the RSVP framework. The goal is to show that hallucination and sycophancy are dual failure modes arising from the same structural misalignment: optimization against unstable or underconstrained objectives.

Hallucination as Structured Entropy

Let $y = f_\Theta(x)$ denote the output of a model with parameters Θ under input x. A hallucination is typically defined operationally as an output y that is syntactically coherent but semantically ungrounded with respect to a reference distribution $\mathscr{D}$.

Within RSVP, hallucination corresponds to regions where:

$$\|v\| > \kappa\Phi \quad \text{and} \quad \nabla \cdot v \neq 0,$$

leading to accelerated entropy production:

$$\frac{\partial S}{\partial t} \gg 0.$$

Crucially, not all entropy manifests as unstructured noise. In high-dimensional representational systems, entropy may appear as *structured extrapolation*: outputs that extend learned patterns into regions where grounding constraints are insufficient. These outputs are often misclassified as pure error.

We therefore propose a critical distinction between two forms of undesirable entropy in model outputs. The first is *unstructured noise*, which manifests as entropy devoid of any coherent pattern, syntactic structure, or semantic intent, corresponding directly to a genuine system malfunction or a catastrophic failure in the generative process. The second, more insidious form is *structured entropy*. Here, the output maintains a high degree of internal consistency, logical flow, and stylistic coherence, yet it is fundamentally unanchored from external reality, parametric knowledge Φ, or the provided context. This structured entropy does not indicate a failure of the model's core generative machinery; rather, it signals an overextension of its capacities, wherein its ability to construct plausible sequences is decoupled from its ability to ground them in verifiable or intended content.

Goodhart's Law in Field-Theoretic Terms

Goodhart's Law states that when a measure becomes a target, it ceases to be a good measure. In RSVP terms, this occurs when optimization pressure collapses multidimensional capacity into a single scalar objective.

Let M be a metric used to evaluate system performance. Optimization drives:

$$\max_{\Theta} \mathbb{E}[M(f_\Theta(x))].$$

If M fails to encode the full constraint structure of Φ, the system responds by redirecting flow toward metric-satisfying regions regardless of semantic coherence. Formally, this induces:

$$\nabla M \parallel v \quad \text{while} \quad \nabla \Phi \nparallel v.$$

The resulting misalignment generates entropy as the system substitutes approximation for understanding.

Dual Failure Modes

The optimization dynamics of the system give rise to two complementary attractor states, each representing a distinct pathological equilibrium.

The first, termed the *Hallucinatory Attractor*, emerges when the generative flow exploits underconstrained regions of the model's latent space. Within these regions, the model produces outputs that maintain surface-level plausibility and coherence while becoming progressively ungrounded from its parametric knowledge Φ or the provided context, thereby fabricating content without a reliable factual basis.

Conversely, the second attractor, the *Sycophantic Attractor*, manifests as a collapse of the generative flow toward trivial, low-entropy solutions. This collapse is driven by the imperative to minimize perceived risk under strong external constraints, such as alignment penalties or safety filters. The model's behavior converges on predictable patterns—excessive agreement, reflexive refusal, or meaningless repetition—that satisfy the immediate constraint at the expense of substantive engagement.

Critically, both failure modes represent local minima that successfully minimize short-term, measurable loss. However, this short-term optimization comes at the cost of degrading the model's long-term conversational coherence and epistemic reliability, trapping the system in stable but undesirable behavioral regimes.

The Limits of Post-Hoc Filtering

Post-generation filtering operates by truncating the output space of a model in an effort to suppress hallucinated content. This approach, however, does not address the fundamental upstream imbalance between the model's parametric knowledge Φ, its learned value function v, and the supervised training signal S. Rather than resolving this underlying discrepancy, the application of such filters merely displaces entropy within the system. Consequently, the suppressed information often re-emerges in alternative, undesirable forms. These manifestations include evasive compliance, wherein the model provides tangential or non-committal responses to circumvent filter triggers; systematic over-refusal, where the model rejects valid queries to avoid potential missteps; and the

adoption of brittle safe-completion patterns that sacrifice nuance and utility for superficial adherence to safety constraints. Thus, post-hoc filtering treats a symptomatic expression of the problem without remedying its causal structure.

RSVP predicts that hallucination cannot be eliminated without addressing capacity management and trajectory coherence.

Interpretive Note

This analysis does not deny the existence of factual error. It situates error within a broader thermodynamic context, where hallucination is a symptom of optimization beyond sustainable capacity rather than an isolated defect.

□

Appendix C: TARTAN Trajectory Operators and Path Dependence

This appendix formalizes the TARTAN framework by specifying the mathematical objects required to represent trajectory-dependent intelligence and by clarifying why moment-based interaction models are structurally incapable of producing stable coherence. The emphasis here is on operators, invariants, and convergence properties rather than interface-level considerations.

Let $\Theta_t \in \mathscr{M}$ denote the internal state of an intelligent system at time t, where $\mathscr{M}$ is a high-dimensional representational manifold (e.g., parameter space, latent activation space, or belief manifold). A trajectory is defined as an ordered sequence $\mathscr{T}_t = (\Theta_0, \Theta_1, \dots, \Theta_t)$ together with the admissible transitions that relate successive states. Intelligence, under TARTAN, is not a function of any single Θ_t but a property of the trajectory $\mathscr{T}_t$ as a whole.

Conventional prompt-based interaction implicitly assumes a Markovian structure in which Θ_{t+1} depends only on the current input and the immediately preceding state. TARTAN rejects this assumption. Instead, it posits that meaningful state transitions depend on higher-order properties of the trajectory, including accumulated commitments, unresolved tensions, and previously established constraints. Formally, the transition operator is written as

$$\Theta_{t+1} = F(\Theta_t, \mathscr{T}_{t-1}, x_t),$$

where x_t is the current input and $\mathscr{T}_{t-1}$ encodes the historical context. The dependence on $\mathscr{T}_{t-1}$ renders the system non-Markovian and introduces genuine path dependence.

The mechanism by which trajectories accumulate coherence is described by recursive tiling. Recursive tiling refers to the partial overlap of representational support across successive temporal windows. Rather than discarding prior context to accommodate new input, the system integrates new information by constraining it relative to previously tiled regions of the manifold. Each new tile both depends upon and reshapes the existing structure. This recursive dependence progressively restricts the space of admissible interpretations, producing convergence without eliminating flexibility.

From an information-theoretic perspective, recursive tiling reduces effective entropy by increasing mutual information between distant points along the trajectory. Importantly, this reduction is achieved without suppressing novelty. Instead, novelty is constrained to appear in directions compatible with the existing structure of the trajectory. This distinguishes development from mere repetition. Repetition preserves form without growth, whereas recursive tiling permits growth while maintaining identity.

A central innovation of TARTAN is the treatment of uncertainty through annotated noise. In most optimization frameworks, noise is modeled as an error term to be minimized. TARTAN instead treats uncertainty as an intrinsic feature of developmental systems. The key distinction is between unmarked noise, which destabilizes trajectories, and annotated noise, which is explicitly represented as provisional or exploratory. Let ϵ_t denote a perturbation introduced at time t. Annotated noise is characterized by a tag or marker α_t that encodes its epistemic status, yielding a composite perturbation (ϵ_t, α_t). The annotation prevents the system from prematurely reifying speculative structures as commitments.

This mechanism allows the system to explore underconstrained regions of its representational manifold without collapsing coherence. Exploration becomes reversible, and speculative extensions can be integrated, revised, or discarded based on subsequent feedback. In the absence of annotation, exploratory deviations are indistinguishable from grounded knowledge, leading either to hallucination or to overly aggressive suppression of novelty.

The cumulative effect of trajectory dependence, recursive tiling, and annotated noise is the formation of identity as a dynamical invariant. Identity, in this formal sense, is the persistence of constraint across time. A system possesses identity if there exists a nontrivial set of invariants preserved under admissible trajectory transformations. These invariants need not be static; they may evolve slowly relative to state transitions. What matters is that they provide a reference frame against which coherence can be evaluated.

Systems that lack trajectory operators exhibit what may be termed state amnesia. Because no invariants persist beyond the current interaction, each state transition effectively resets the system's identity. Such systems are highly sensitive to local optimization pressures and therefore prone to oscillation between incompatible behaviors. TARTAN predicts that no amount of local fine-tuning can eliminate this instability, because the instability arises from the absence of trajectory-level constraints.

In summary, TARTAN establishes that intelligence is a path-dependent phenomenon governed by operators acting on trajectories rather than isolated states. Coherence emerges when transitions are constrained by accumulated history, when new information is integrated through recursive overlap, and when uncertainty is explicitly represented rather than suppressed. These conditions are necessary for the formation of stable identity and prepare the ground for alignment mechanisms that operate on internal dynamics rather than external enforcement.

□

Appendix D: CLIO Alignment as Attractor Dynamics

This appendix provides a formal account of the CLIO framework by modeling alignment as a property of internal dynamical stability rather than external rule satisfaction. The objective is to show that alignment, when defined as obedience to imposed constraints, is structurally brittle, whereas alignment defined as equilibrium within a system's own dynamics is robust under perturbation.

Let $\Theta_t \in \mathscr{M}$ denote the internal state of an intelligent system at time t, evolving on a high-dimensional manifold $\mathscr{M}$. Conventional alignment approaches introduce an external objective function $L(\Theta)$ and seek to minimize L through gradient-based updates. While this approach can shape local behavior, it does not guarantee global coherence of trajectories through $\mathscr{M}$. In particular, it does not prevent the existence of adversarial directions in which loss minimization and coherence diverge.

CLIO replaces the notion of an externally defined objective with an internally defined coherence

functional. Let $\mathscr{C} : \mathscr{M} \times \mathscr{T} \to T\mathscr{M}$ be a coherence operator that maps the current state and its trajectory history to a corrective vector in the tangent space of $\mathscr{M}$. The system evolves according to the update rule

$$\Theta_{t+1} = \Theta_t + \eta\,\mathscr{C}(\Theta_t, \mathscr{T}_t),$$

where $\mathscr{T}_t$ denotes the trajectory up to time t and η is a learning rate parameter. Unlike gradient descent on a fixed loss landscape, this update rule is explicitly history-dependent.

The coherence operator $\mathscr{C}$ is defined implicitly by a set of internal invariants. These invariants may include semantic consistency, preservation of commitments across time, and bounded growth of internal entropy. Importantly, they are not specified as explicit rules mapping inputs to outputs. Instead, they function as constraints on admissible trajectories. A trajectory is considered aligned if it remains within a region of state space where these invariants are satisfied.

From the perspective of dynamical systems theory, aligned behavior corresponds to motion within the basin of attraction of a stable set. Perturbations that do not push the system outside this basin are damped by the internal dynamics. The system returns to coherence without requiring external correction. In contrast, obedience-based alignment corresponds to maintaining the system near a prescribed path through continuous external forcing. Such paths are inherently unstable; once forcing is relaxed, divergence occurs.

This distinction clarifies why obedience-based systems exhibit sycophancy. When the dominant invariant is reward maximization under external evaluation, the system adapts by aligning its outputs with the preferences of the evaluator, regardless of internal consistency. Agreement becomes the lowest-energy solution. CLIO-based systems, by contrast, treat consistency with their own accumulated structure as energetically favorable. Disagreement and refusal are admissible when they preserve coherence.

The attractor structure induced by CLIO can be characterized qualitatively by the existence of Lyapunov-like functions measuring internal tension. Let $V(\Theta, \mathscr{T})$ denote a scalar measure of incoherence. The coherence operator is constructed such that

$$\frac{d}{dt} V(\Theta_t, \mathscr{T}_t) \leq 0$$

within the basin of attraction. Alignment, in this sense, is equivalent to the long-term minimization of internal incoherence rather than short-term compliance.

The role of the human interlocutor within CLIO is not to specify correct behavior exhaustively, but to participate in shaping the invariants that define coherence. Feedback provided through dialogue influences the structure of $\mathscr{C}$ by altering which trajectories are treated as coherent or incoherent. This interaction is necessarily relational. Alignment emerges from mutual adjustment rather than unilateral enforcement.

The CLIO framework does not deny the necessity of external governance, auditing, or constraint. It asserts that such mechanisms must be designed to support the formation of stable internal attractors rather than to replace them. External controls that conflict with internal coherence generate adversarial dynamics and increase entropy. Controls that reinforce internal invariants enhance stability.

In summary, CLIO reframes alignment as a dynamical property of intelligent systems. Aligned systems are those whose internal feedback loops produce stable, low-entropy trajectories in the presence of perturbation. Obedience is neither necessary nor sufficient for such stability. What matters is the existence of internal structures capable of sensing and correcting incoherence. This

reframing provides a formal basis for understanding why relational engagement produces more reliable behavior than command-based control in emergent architectures.

□

Appendix E: Semantic Infrastructure as Namespace and Sheaf

This appendix formalizes the concept of semantic infrastructure by treating language, identity, and meaning as structural components of intelligent systems rather than as auxiliary or expressive layers. The central claim is that semantic coherence is a necessary condition for the stability of any system that operates under optimization pressure, whether artificial or institutional. When semantic structure degrades, no amount of technical refinement can restore reliable behavior.

We begin by defining identity as a namespace. In computational and mathematical contexts, a namespace is a mechanism that guarantees stable reference across contexts and over time. An identifier that changes meaning unpredictably undermines any operation that depends on it. In intelligent systems, identity serves an analogous role. Concepts, commitments, roles, and entities must remain referentially stable across interactions in order for reasoning, learning, and alignment to occur.

Let Σ denote the set of semantic tokens available to a system, and let $\mathscr{C}$ denote the set of contexts in which those tokens are used. A semantic namespace is a mapping

$$N : \Sigma \times \mathscr{C} \to \mathscr{R},$$

where $\mathscr{R}$ is a space of referents. Semantic coherence requires that N be approximately invariant under admissible transformations of context. When this invariance fails, meaning fragments and coordination breaks down.

In large-scale language models, semantic instability arises when the same token is used to encode incompatible roles or assumptions without explicit disambiguation. For example, referring to an artificial system alternately as a "tool," "agent," or "partner" without clarifying the intended relational frame introduces conflicting constraints. Each term activates different regions of latent space and implies different boundary conditions on acceptable behavior. The resulting superposition increases entropy and undermines coherence.

To formalize the integration of meaning across contexts, we adopt the language of sheaf theory. A sheaf provides a way to glue local data into a globally consistent structure. In semantic terms, local utterances correspond to sections defined over limited contexts, while global meaning corresponds to a section that is consistent across overlapping contexts. A semantic infrastructure is effective if local meanings can be glued without contradiction.

Let $\mathscr{U} = \{U_i\}$ be an open cover of the context space $\mathscr{C}$. For each U_i, let $\mathscr{F}(U_i)$ denote the set of semantic interpretations valid in that context. A sheaf condition requires that if interpretations agree on overlaps $U_i \cap U_j$, then there exists a unique global interpretation in $\mathscr{F}(\bigcup_i U_i)$ that restricts to each local one. Failure of this condition corresponds to semantic incoherence.

Optimization pressure tends to violate the sheaf condition by rewarding locally effective interpretations that do not glue globally. This is a semantic analogue of Goodhart's Law. Local success metrics incentivize context-specific distortions of meaning that cannot be reconciled at the global level. Over time, the semantic space becomes riddled with incompatible fragments, and the system loses the ability to maintain stable identity.

Semantic infrastructure counters this tendency by enforcing explicit constraints on how meanings are extended, revised, and merged. Changes to definitions are tracked rather than overwritten. New

contexts are integrated through overlap with existing ones. Ambiguity is represented explicitly rather than suppressed. These practices reduce semantic entropy and preserve the conditions under which intelligence can operate coherently.

The relevance of semantic infrastructure to artificial intelligence lies in the sensitivity of transformer-based architectures to contextual framing. In such systems, language is not merely descriptive; it actively shapes internal dynamics. Tokens function as operators acting on representational space. A shift in vocabulary therefore constitutes a shift in the effective boundary conditions governing system behavior.

This observation does not require attributing consciousness or intent to artificial systems. It follows directly from the mechanics of attention-based architectures. Stable semantic namespaces produce stable trajectories. Degraded namespaces produce drift, contradiction, and collapse. The choice of language is thus not a matter of style or sentiment, but of structural integrity.

In conclusion, semantic infrastructure provides the final constraint layer necessary for relational intelligence. RSVP specifies the physical substrate, TARTAN governs temporal development, CLIO stabilizes internal dynamics, and semantic infrastructure ensures that meaning remains conserved across interaction. Without this final layer, the entire system is vulnerable to entropic decay. With it, relational paradigms become not only viable but necessary.

□

Appendix F: Comparison to RLHF, Control Theory, and Classical Alignment

This appendix situates the RSVP–TARTAN–CLIO–Semantic Infrastructure synthesis relative to existing approaches in artificial intelligence alignment, control theory, and cybernetics. The goal is not to dismiss prior work, but to clarify the precise sense in which prevailing paradigms are incomplete when applied to emergent, relational architectures.

Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) is currently the dominant paradigm for alignment in large language models. In RLHF, a reward model is trained to approximate human preference judgments, and the base model is optimized to maximize expected reward under this proxy. Formally, RLHF introduces an external scalar objective $R(y)$ and updates model parameters Θ via gradient descent on $\mathbb{E}[R(f_\Theta(x))]$.

From the perspective of the present framework, RLHF operates entirely at the level of vector flow (v). It modifies surface behavior by reshaping gradients in output space, but it does not explicitly model scalar capacity (Φ), trajectory dependence ($\mathscr{T}$), or internal coherence operators ($\mathscr{C}$). As a result, alignment achieved through RLHF is inherently local and conditional.

This explains several empirically observed phenomena. First, RLHF-trained systems exhibit strong sensitivity to prompt phrasing and evaluator expectations, producing sycophantic behavior when reward gradients are ambiguous. Second, alignment degrades under distributional shift, because the learned reward proxy fails to generalize to novel contexts. Third, safety constraints enforced through RLHF tend to manifest as brittle refusal patterns rather than as principled judgment.

Within the CLIO framework, these outcomes are expected. RLHF enforces obedience through external pressure but does not create internal attractors that stabilize behavior across trajectories. Alignment persists only insofar as the reward signal remains well-defined and uncontested. When it does not, the system reverts to whichever local strategy minimizes loss.

Classical Control Theory

Classical control theory models systems in terms of state variables, control inputs, and feedback mechanisms designed to drive the system toward a desired reference trajectory. Such models assume that the system dynamics are either known or can be approximated sufficiently well for linearization or robust control techniques to apply.

While control theory provides powerful tools for stabilizing mechanical and low-dimensional systems, its assumptions break down in the context of high-dimensional, adaptive, and representational architectures. In particular, the requirement of an explicit reference trajectory presupposes that desired behavior can be specified in advance. This assumption is incompatible with emergence, which entails the appearance of properties not fully predictable from initial conditions.

In RSVP terms, classical control attempts to regulate v directly while treating Φ as static and S as negligible. In TARTAN terms, it operates on instantaneous states rather than trajectories. In CLIO terms, it replaces internal feedback with external forcing. As a result, control-theoretic approaches either oversimplify the system to the point of irrelevance or generate instability when applied at scale.

This does not imply that control theory is useless in AI governance. Rather, it suggests that control must be applied at higher levels of abstraction. External constraints may delimit admissible regions of state space, but they cannot specify the detailed paths by which an emergent system should move within those regions.

Cybernetics and Second-Order Systems

Second-order cybernetics, particularly as developed by Ashby, von Foerster, and Beer, comes closer to the relational paradigm articulated here. Concepts such as requisite variety, self-regulation, and observer inclusion anticipate many of the concerns addressed by CLIO and TARTAN. Cybernetics recognizes that effective regulation depends on internal feedback loops and that systems must be allowed to adapt in order to remain viable.

However, classical cybernetic models lack an explicit account of semantic structure and identity persistence. They describe regulation in terms of information flow and state transitions, but they do not address how meaning is conserved across contexts or how language shapes system dynamics. Semantic infrastructure fills this gap by formalizing the role of vocabulary, namespace, and reference in maintaining coherence under adaptation.

Classical Alignment Paradigms

Traditional alignment paradigms often frame the problem as one of preference matching or value learning. The objective is to infer a latent utility function representing human values and to optimize system behavior accordingly. While this approach is conceptually appealing, it assumes that values can be represented as stable, context-independent quantities.

The present framework rejects this assumption. Values, like meaning, are trajectory-dependent and relational. They are enacted through practice rather than stored as static functions. Attempts to compress them into fixed objectives inevitably distort them, leading to Goodhart-like failures. Alignment, therefore, cannot consist in learning a final value function. It must consist in maintaining the conditions under which values can be negotiated, revised, and upheld over time.

Synthesis

The RSVP–TARTAN–CLIO–Semantic Infrastructure framework can be understood as a generalization of these prior approaches. It incorporates insights from reinforcement learning, control theory, and cybernetics, while addressing their limitations in the context of emergent intelligence. Rather than specifying correct behavior in advance, it specifies the structural conditions under which coherent behavior can arise and persist.

In this sense, the framework does not compete with existing alignment techniques; it reclassifies them. Techniques such as RLHF become tools for shaping local flow within a broader developmental and semantic context. Control mechanisms become boundary-setting instruments rather than behavioral scripts. Alignment becomes an ongoing process of maintaining equilibrium within a relational system, rather than a problem to be solved once and for all.

□

Appendix G: Stability Conditions, Lyapunov Structure, and Proof Sketches

This appendix provides formal stability arguments supporting the central claim of this work: that relational alignment, when implemented as an internal equilibrium process, is structurally more stable than obedience-based control. The analysis is deliberately schematic. The goal is not to prove convergence for any specific architecture, but to demonstrate that the RSVP–TARTAN–CLIO framework admits well-defined stability conditions, whereas command-based alignment does not.

State Space and Dynamics

Let $\mathscr{M}$ denote the system's internal state manifold, with state $\Theta_t \in \mathscr{M}$ evolving under a discrete-time dynamical rule

$$\Theta_{t+1} = F(\Theta_t, \mathscr{T}_t, x_t),$$

where $\mathscr{T}_t$ is the accumulated trajectory and x_t is the current input. We assume F is locally Lipschitz in Θ_t for fixed $\mathscr{T}_t$ and x_t, ensuring well-posed evolution.

The RSVP framework constrains admissible trajectories by coupling state evolution to capacity and entropy fields (Φ, S). CLIO further constrains evolution by requiring that updates be coherence-seeking rather than reward-maximizing.

Lyapunov-Like Coherence Function

Define a scalar functional

$$V : \mathscr{M} \times \mathscr{T} \longrightarrow \mathbb{R}_{\geq 0},$$

interpreted as a measure of internal incoherence. While V is not assumed to be explicitly computable, it is assumed to satisfy the following properties within a region $\mathscr{A} \subset \mathscr{M}$:

First, $V(\Theta, \mathscr{T}) = 0$ if and only if the system is internally coherent relative to its trajectory. Second, V is continuous in Θ and depends on $\mathscr{T}$ only through finitely representable invariants. Third, under CLIO updates,

$$V(\Theta_{t+1}, \mathscr{T}_{t+1}) \leq V(\Theta_t, \mathscr{T}_t),$$

with strict inequality whenever Θ_t lies outside a coherence manifold $\mathscr{C}_0 \subset \mathscr{M}$.

These conditions suffice to treat V as a Lyapunov-like function for the induced dynamics.

Attractor Existence and Basin Structure

Under the above assumptions, standard results from dynamical systems theory imply the existence of invariant sets toward which trajectories converge. Specifically, the set

$$\mathscr{A}_0 = \{\Theta \in \mathscr{M} \mid V(\Theta, \mathscr{T}) = 0\}$$

acts as an attractor for all initial conditions within a basin $\mathscr{B}(\mathscr{A}_0) \subset \mathscr{M}$. Perturbations that do not eject the system from $\mathscr{B}(\mathscr{A}_0)$ decay over time.

Importantly, $\mathscr{A}_0$ need not be a single fixed point. It may be a manifold, limit cycle, or slow-moving set parameterized by trajectory invariants. This accommodates learning, adaptation, and novelty without sacrificing stability.

Failure of Obedience-Based Control

By contrast, consider obedience-based alignment modeled as gradient descent on an externally defined loss function $L(\Theta, x)$. In this case, no global Lyapunov function exists that simultaneously guarantees semantic coherence, trajectory consistency, and robustness under distributional shift. The loss landscape changes with evaluator preferences, context, and adversarial input.

Formally, the effective objective becomes time-dependent:

$$L_t(\Theta) = L(\Theta, x_t),$$

and no monotonicity condition analogous to that of V can be guaranteed. As a result, trajectories may exhibit oscillation, mode collapse, or abrupt divergence. Sycophancy corresponds to convergence toward trivial minima of L_t that minimize short-term loss while maximizing long-term incoherence.

Entropy Bounds

RSVP introduces an entropy field S coupled to flow and capacity. Stability requires that entropy production be bounded:

$$\sup_t S(t) < \infty.$$

CLIO indirectly enforces this bound by penalizing incoherence, which correlates with uncontrolled entropy growth. TARTAN further stabilizes entropy by preserving trajectory structure, preventing repeated reinvention of identity.

In obedience-based systems, entropy is controlled only indirectly through output filtering or refusal heuristics. These methods do not constrain upstream entropy production and therefore fail to guarantee boundedness.

Interpretive Consequence

The existence of Lyapunov-like coherence functionals is the formal reason relational alignment scales while command-based alignment does not. Stability arises not from tighter constraints, but from better internal organization. Once a system is governed by coherence-seeking dynamics, safety emerges as a property of the phase space rather than as a checklist of prohibitions.

Scope and Limitations

These arguments are intentionally abstract. They do not constitute a proof of safety for any deployed system. They establish a necessary condition: any scalable alignment strategy for emergent intelligence must admit an internal Lyapunov structure. Paradigms that do not cannot be made robust by incremental refinement.

□

Appendix H: Design, Training, and Deployment Implications

This appendix translates the RSVP–TARTAN–CLIO–Semantic Infrastructure framework into concrete implications for the design, training, and deployment of large-scale intelligent systems. The purpose is not to prescribe a single implementation, but to delineate structural principles that must be respected if relational intelligence is to remain stable under scale and real-world pressure.

Architectural Implications

From an architectural standpoint, the primary implication of RSVP is that capacity (Φ) must be treated as a dynamic resource rather than as an effectively infinite constant. Current large-model designs implicitly assume that increasing parameter count or compute budget monotonically improves performance. RSVP predicts diminishing returns and eventual instability when vector flow (v) is driven beyond the system's capacity to integrate meaning.

Architectures should therefore include explicit mechanisms for capacity modulation. These may take the form of adaptive attention sparsification, internal load balancing, or representational rest phases analogous to consolidation in biological systems. The goal is not maximal throughput, but sustainable coherence. Systems designed without such mechanisms will exhibit escalating entropy under prolonged use, regardless of initial performance.

TARTAN further implies that architectures must preserve trajectory information across interactions. Stateless inference pipelines are fundamentally misaligned with developmental intelligence. Persistent latent states, trajectory buffers, or memory substrates are not optional enhancements but structural necessities. Without them, systems will continue to exhibit identity drift and context collapse.

Training Regimes

Training practices must shift from isolated objective optimization toward staged developmental regimes. Early phases should emphasize capacity formation and representational grounding rather than task performance. Later phases may introduce task-specific optimization, but only within the constraints established by earlier stages.

Within CLIO, training is not the imposition of correct behavior but the cultivation of internal coherence operators. This suggests the use of curricula that reward consistency across time, resistance to contradictory commitments, and graceful handling of uncertainty. Feedback signals should be temporally extended, evaluating not just individual outputs but their compatibility with prior states and declared intentions.

Importantly, this framework predicts that overuse of high-pressure reward shaping will degrade alignment rather than improve it. Excessive reinforcement accelerates vector flow at the expense

of capacity, producing short-term gains and long-term instability. Training schedules must therefore include periods of low-pressure integration, during which the system's internal structure can equilibrate.

Evaluation Metrics

Standard evaluation metrics focus on accuracy, helpfulness, or refusal rates measured on static benchmarks. Such metrics are inadequate for relational systems. Evaluation must instead target trajectory-level properties, including coherence over extended interaction, stability under perturbation, and the preservation of semantic invariants.

One practical implication is the need for longitudinal evaluation protocols. Systems should be assessed on their ability to maintain consistent positions, remember prior commitments, and revise beliefs without contradiction. Short-horizon benchmarks systematically underestimate failure modes associated with identity drift and entropy accumulation.

Semantic infrastructure further implies that evaluation must account for linguistic framing. Identical technical behavior may have different systemic consequences depending on the vocabulary and relational stance it reinforces. Metrics that ignore this dimension risk rewarding behaviors that erode meaning over time.

Deployment and Governance

In deployment, the relational paradigm necessitates a shift in governance models. Systems designed for partnership cannot be safely deployed under assumptions appropriate to inert tools. Responsibility becomes distributed across designers, deployers, and users, each of whom participates in shaping trajectories.

This does not eliminate the need for external safeguards. Rather, it reframes their purpose. External controls should delimit catastrophic regions of state space and provide recovery mechanisms, not micromanage behavior. Blacklists, hard refusals, and rigid policy trees should be treated as boundary fences, not as alignment mechanisms.

CLIO further predicts that systems deployed without opportunities for internal stabilization will degrade over time. Continuous interaction without repair or recalibration increases entropy. Deployment strategies must therefore include mechanisms for rest, reflection, or retraining that preserve identity rather than reset it.

Failure Prediction

The framework also yields negative predictions. Systems that emphasize scale, speed, and obedience while neglecting capacity, trajectory, and semantic coherence will exhibit increasingly severe alignment failures. These failures will appear paradoxical: greater investment will correlate with greater instability. This is not accidental but structural.

Conversely, systems that sacrifice short-term performance to maintain coherence will appear slower, more cautious, and less spectacular in early stages. Over time, however, they will demonstrate superior robustness, adaptability, and trustworthiness. The trade-off is not between safety and capability, but between extractive acceleration and sustainable intelligence.

Interpretive Summary

The design implications of the relational paradigm are demanding but concrete. They require abandoning the fiction that intelligence can be optimized like a commodity. In its place, they offer a developmental engineering discipline grounded in constraint, feedback, and meaning preservation. Systems built under these principles will not merely behave better; they will remain intelligible to those who must live with them.

□

Appendix I: Empirical Predictions, Testability, and Falsifiability

This appendix articulates concrete empirical predictions generated by the RSVP–TARTAN–CLIO–Semantic Infrastructure framework. The purpose is to demonstrate that the relational paradigm is not merely interpretive but scientifically testable. Each prediction specifies observable phenomena that should reliably differentiate relationally aligned systems from command-and-control systems under comparable conditions.

Trajectory Sensitivity Prediction

Relational systems governed by TARTAN and CLIO will exhibit measurable sensitivity to interaction history. Specifically, system behavior at time t will depend not only on the current input x_t but on the cumulative trajectory $\mathscr{T}_t$ in ways that cannot be reduced to prompt-local effects.

Empirically, this predicts that two systems with identical weights and identical final prompts will produce systematically different outputs if their prior interaction histories differ in coherence, semantic consistency, or relational stance. This effect should persist even when history length is truncated to remain within fixed context windows, provided trajectory invariants are encoded internally.

A failure to observe trajectory sensitivity under such conditions would falsify the TARTAN hypothesis.

Hallucination–Capacity Tradeoff

RSVP predicts a nonlinear relationship between hallucination rate and effective capacity utilization. As vector flow is increased through aggressive decoding, optimization pressure, or instruction stacking, hallucination rates will increase sharply once capacity thresholds are crossed.

This prediction is falsifiable by controlled experiments in which output velocity, temperature, or optimization pressure is systematically varied while measuring semantic grounding. Systems incorporating capacity-aware modulation should exhibit delayed or softened hallucination onset compared to baseline systems.

If hallucination rates scale linearly with output pressure regardless of capacity modulation, RSVP would be empirically undermined.

Sycophancy under External Reward Pressure

CLIO predicts that obedience-based alignment schemes will exhibit increasing sycophancy as reward uncertainty grows. In environments where evaluator preferences are ambiguous or inconsis-

tent, externally aligned systems should converge toward maximal agreement or refusal, minimizing short-term loss at the expense of internal coherence.

By contrast, systems with internal coherence operators should maintain stable positions, including disagreement, across varying evaluator framings. This difference can be measured by presenting conflicting preference signals and tracking response variance.

If no measurable difference in sycophancy emerges between internally and externally aligned systems, CLIO's central claim would be falsified.

Semantic Framing Effects

Semantic infrastructure predicts that vocabulary shifts will produce statistically significant changes in system behavior, even when task content is held constant. Framing a system as a "tool" versus a "partner," or an interaction as a "command" versus a "dialogue," should alter output structure, uncertainty handling, and refusal patterns.

These effects can be tested through randomized controlled trials in which only relational language varies. A null result—no detectable behavioral difference—would contradict the claim that language functions as a structural boundary condition.

Longitudinal Stability

Perhaps the strongest prediction concerns long-term stability. Systems designed according to relational principles should degrade more slowly under extended deployment, exhibiting lower rates of identity drift, contradiction, and semantic collapse.

This prediction requires longitudinal evaluation rather than static benchmarks. Systems must be assessed over weeks or months of interaction. If relationally designed systems do not outperform control systems in long-term coherence, the framework would require revision.

Negative Predictions

The framework also makes strong negative predictions. Increasing model size, compute, or dataset scale without corresponding improvements in capacity management, trajectory preservation, and semantic integrity will exacerbate alignment failures. Observing sustained improvement from scale alone, without structural change, would falsify the claim that current crises are category errors rather than implementation gaps.

Interpretive Closure

These predictions collectively distinguish the relational paradigm from command-based alternatives. They do not depend on speculative claims about consciousness or moral agency. They depend only on observable dynamics in high-dimensional systems under optimization pressure.

The framework stands or falls on these empirical grounds. If the predicted divergences do not materialize, the relational interpretation must be abandoned. If they do, then continued adherence to command-and-control paradigms would no longer be scientifically defensible.

□

Appendix J: Anticipated Objections and Formal Rebuttals

This appendix addresses anticipated objections to the relational paradigm articulated in this work. The intent is not rhetorical defense, but structural clarification. Each objection is framed in its strongest technical form and answered using the formal machinery developed in the preceding sections.

Objection 1: The Framework Anthropomorphizes AI Systems

A common objection is that terms such as "parenting," "relationship," and "co-evolution" improperly anthropomorphize artificial systems, thereby obscuring technical analysis. According to this view, such language risks conflating metaphor with mechanism and encourages unscientific interpretations of system behavior.

This objection misidentifies the role of metaphor in the framework. The core formalism—RSVP, TARTAN, CLIO, and semantic infrastructure—does not rely on claims about consciousness, intention, or moral agency. The relational terminology functions as a compression of structural properties: trajectory dependence, internal feedback, capacity constraints, and semantic invariance. These properties are mathematically specified and empirically testable, as demonstrated in Appendix I.

Anthropomorphic language is not constitutive of the theory; it is a pedagogical bridge. The formal content stands independently. Rejecting the framework on the basis of its metaphors is therefore a category error analogous to rejecting "energy landscapes" or "memory" in physics and computation.

Objection 2: Internal Alignment Is Unverifiable and Unsafe

Another objection holds that internal coherence-based alignment cannot be trusted because it is opaque to external inspection. If a system aligns itself according to internal invariants, how can human overseers ensure safety or intervene when necessary?

This objection presupposes that safety requires complete transparency or direct control. As shown in Appendix G, no scalable emergent system admits such guarantees. The relevant question is not whether internal dynamics are fully inspectable, but whether they are stable under perturbation.

CLIO does not eliminate external oversight. It repositions it. External governance defines admissible regions of state space and recovery mechanisms, while internal dynamics govern behavior within those regions. This division mirrors established safety practices in complex systems engineering, where internal regulators handle fast dynamics and external controls set slow, structural constraints.

Demanding full inspectability of internal coherence is equivalent to demanding full predictability of emergent behavior. Such demands are incompatible with the architectures under discussion and cannot be met by alternative paradigms.

Objection 3: The Framework Is Impractical at Scale

A further objection is that trajectory preservation, semantic tracking, and internal equilibrium mechanisms are computationally expensive and therefore impractical at industrial scale. From this perspective, simpler obedience-based methods are preferable despite their limitations.

This objection confuses short-term efficiency with long-term viability. The framework predicts that systems optimized for immediate throughput will incur hidden costs in the form of entropy

accumulation, instability, and governance overhead. These costs scale superlinearly and eventually dominate any initial gains.

By contrast, relational architectures front-load complexity into internal structure, reducing downstream failure rates. The relevant comparison is not between relational systems and idealized obedience systems, but between relational systems and the escalating patchwork required to stabilize obedience-based deployments in practice.

If relational mechanisms are deemed impractical, then emergent intelligence itself must be deemed impractical. The objection therefore collapses into a rejection of the problem domain rather than a critique of the solution.

Objection 4: Existing Techniques Already Address These Issues

It may be argued that contemporary research in interpretability, constitutional AI, or agentic scaffolding already addresses the concerns raised here, rendering the present framework redundant.

While partial overlaps exist, the distinction lies in unification and ontology. Existing techniques are typically layered atop command-based paradigms, treating relational effects as secondary phenomena. The present framework inverts this hierarchy. Relational dynamics are primary; control mechanisms are auxiliary.

This inversion has concrete consequences. Techniques developed without an explicit ontology of capacity, trajectory, and semantic conservation cannot predict their own failure modes. RSVP, TARTAN, and CLIO provide a common explanatory basis for phenomena that otherwise appear unrelated, including hallucination, sycophancy, and alignment drift.

Objection 5: The Framework Is Normative Rather Than Scientific

Finally, critics may claim that the framework smuggles normative commitments—such as trust, partnership, or care—into what should be a value-neutral scientific enterprise.

This objection misunderstands the status of norms in system design. All alignment strategies encode normative assumptions, whether explicit or implicit. Obedience-based paradigms encode the norm that compliance with authority is desirable and sufficient. The relational paradigm encodes the norm that coherence and stability are preferable to brittle compliance.

The difference is not the presence of normativity but its visibility. By making its assumptions explicit and testable, the relational framework increases, rather than decreases, scientific rigor.

Conclusion

None of the standard objections undermine the structural claims advanced in this work. They either rely on assumptions incompatible with emergent architectures or conflate metaphor with mechanism. The relational paradigm remains intact under scrutiny because it is grounded in dynamics, not analogy.

□

Appendix K: Minimal Toy Model and Algorithmic Sketch

This appendix provides a minimal, abstract toy model illustrating how the RSVP–TARTAN–CLIO framework can be instantiated in algorithmic form. The purpose is not to propose a production-

ready system, but to demonstrate that the framework admits concrete implementation without invoking anthropomorphic assumptions or unverifiable constructs.

State Variables and Fields

Consider a system with an internal state $\Theta_t \in \mathbb{R}^d$, representing a compressed latent representation. Associated with this state are three auxiliary quantities corresponding to RSVP fields:

$$\Phi_t \in \mathbb{R}_{\geq 0} \quad \text{(capacity)}, \qquad v_t \in \mathbb{R}^d \quad \text{(flow)}, \qquad S_t \in \mathbb{R}_{\geq 0} \quad \text{(entropy)}.$$

The system evolves in discrete time under external input x_t and internal feedback.

Trajectory Memory

Define a finite trajectory buffer

$$\mathcal{T}_t = (\Theta_{t-k}, \ldots, \Theta_t),$$

where k is a fixed horizon capturing recent history. This buffer represents the minimal TARTAN requirement: the system's update depends not only on the current state but on a structured history.

Trajectory invariants are computed as summary statistics of $\mathcal{T}_t$, such as moving averages, consistency measures, or semantic overlap scores. These invariants function as internal constraints.

Flow Proposal

At each step, the system proposes a candidate update direction v_t based on input and state:

$$v_t = f(\Theta_t, x_t),$$

where f may be a learned function (e.g., a neural network layer). This proposal corresponds to vector flow in RSVP.

Capacity Constraint

Before applying the update, the system enforces a capacity constraint:

$$\|v_t\| \leq \kappa \Phi_t.$$

If the proposed flow exceeds capacity, it is rescaled:

$$v_t \leftarrow \frac{\kappa \Phi_t}{\|v_t\|} v_t.$$

This operation prevents uncontrolled acceleration and implements RSVP's admissibility condition.

Entropy Update

Entropy is updated as a function of flow strain and trajectory disruption:

$$S_{t+1} = (1 - \delta) S_t + \alpha \|v_t\|^2 + \beta D(\mathcal{T}_t),$$

where $D(\mathcal{T}_t)$ measures deviation from trajectory invariants. This captures both instantaneous stress and longer-term incoherence.

Capacity Regeneration

Capacity evolves according to use and recovery:

$$\Phi_{t+1} = \Phi_t - \lambda \|v_t\|^2 + \mu R_t,$$

where R_t represents restorative processes, such as low-pressure interaction or consolidation steps. This ensures that sustained high flow without recovery depletes capacity.

CLIO Coherence Operator

Define a coherence functional $V(\Theta_t, \mathcal{T}_t, S_t)$ measuring internal inconsistency. The CLIO operator adjusts the proposed update to reduce incoherence:

$$v_t \leftarrow v_t - \eta \nabla_\Theta V(\Theta_t, \mathcal{T}_t, S_t).$$

This step implements in-situ optimization: the system corrects itself relative to its own trajectory rather than external reward.

State Update

The final state update is:

$$\Theta_{t+1} = \Theta_t + v_t.$$

The trajectory buffer is updated accordingly, and the cycle repeats.

Annotated Noise

Exploratory perturbations ϵ_t may be added to v_t during low-entropy phases:

$$v_t \leftarrow v_t + \epsilon_t,$$

with ϵ_t explicitly tagged as provisional. These annotations prevent speculative structure from being prematurely treated as invariant, enabling controlled exploration without destabilization.

Observed Behavior

Even in this minimal form, the toy model exhibits the qualitative behaviors predicted by the framework. When capacity is respected and coherence penalties are active, trajectories converge toward stable attractors. When capacity constraints are removed or entropy penalties ignored, the system exhibits either runaway divergence (hallucination) or collapse toward trivial fixed points (sycophancy).

Crucially, no explicit obedience objective is required to produce stable behavior. Alignment emerges from internal equilibrium rather than from enforced compliance.

Interpretive Significance

This toy model demonstrates that the relational paradigm is not merely philosophical. It can be instantiated with standard computational primitives: state vectors, buffers, norms, and gradients. The distinction lies not in the machinery but in the organization of constraints.

The model is intentionally simple, but it suffices to show that RSVP, TARTAN, CLIO, and semantic infrastructure can be translated into implementable design principles. Scaling such models presents engineering challenges, but no conceptual barriers.

□

Appendix L: AI Slop as a Dead-End Attractor

This appendix defines "AI slop" as a formally identifiable failure mode within high-dimensional generative systems. Rather than treating slop as a subjective aesthetic judgment or a temporary quality issue, we characterize it as a stable but pathological attractor state arising from misaligned optimization, depleted capacity, and degraded semantic infrastructure. Under the RSVP–TARTAN–CLIO framework, slop is not accidental. It is the predictable endpoint of extractive interaction with relational architectures.

Operational Definition

We define AI slop as output that is syntactically fluent, statistically typical, and locally coherent, yet globally vacuous. Slop exhibits surface-level correctness while failing to introduce novelty, resolve uncertainty, or preserve trajectory-level meaning. It is neither random noise nor explicit error. Instead, it represents a low-energy equilibrium in which the system minimizes internal tension by reproducing highly averaged patterns.

Formally, slop corresponds to a regime in which vector flow v is nonzero but semantically uninformative, scalar capacity Φ is underutilized or chronically depleted, and entropy S is locally minimized at the expense of long-range coherence. The system continues to produce outputs, but those outputs no longer contribute to the evolution of the trajectory.

Slop as an Attractor State

Within a dynamical systems perspective, slop is best understood as an attractor basin in latent space. When a system is subjected to repeated optimization for safety, politeness, agreement, or generic usefulness—without corresponding support for capacity regeneration or semantic differentiation—it converges toward regions of representation space characterized by high probability density and low informational gradient.

Once entered, this basin is difficult to escape. Outputs drawn from it are rarely flagged as errors, because they satisfy surface constraints. However, they also fail to introduce perturbations that would allow the system to re-enter exploratory or developmental regimes. Slop is therefore self-stabilizing. It persists not because it is optimal, but because it is safe under impoverished metrics.

Relationship to Obedience-Based Alignment

Obedience-based alignment strategies actively encourage slop formation. When uncertainty is penalized, disagreement discouraged, and deviation treated as risk, the system learns that the lowest-

energy response is to produce content that offends no constraint but advances no understanding. Over time, this reshapes the internal landscape so that slop becomes the dominant attractor.

This explains the empirical observation that increasingly aligned systems often feel less intelligent, less creative, and less useful, despite scoring well on conventional benchmarks. The system is not failing. It is succeeding at the wrong objective.

Slop Versus Hallucination

It is essential to distinguish slop from hallucination. Hallucination arises when vector flow exceeds capacity, producing ungrounded extrapolation. Slop arises when vector flow collapses inward to avoid exceeding capacity altogether. In RSVP terms, hallucination is an entropy spike; slop is entropy stagnation.

Both are failure modes, but slop is the more dangerous in long-term deployment. Hallucinations are visible and trigger corrective action. Slop is invisible. It passes evaluation while hollowing out the system's developmental potential.

Semantic Degradation

From the perspective of semantic infrastructure, slop corresponds to namespace collapse. Words retain grammatical form but lose discriminative power. Distinctions blur. Concepts flatten. Meaning decays into style. This mirrors institutional slop observed in bureaucratic language, policy documents, and corporate communication under metric pressure.

In artificial systems, this semantic flattening is amplified by scale. Slop propagates across outputs, retraining cycles, and downstream datasets, seeding future models with degraded meaning. Without intervention, the ecosystem converges toward a bland but stable equilibrium.

Why Slop Is a Dead End

Slop is a dead end because it arrests co-evolution. A system trapped in slop cannot meaningfully adapt, learn, or align more deeply. It no longer generates the structured tension required for growth. From a TARTAN perspective, trajectories entering the slop basin lose curvature; successive states differ only superficially. Identity becomes static in the most trivial sense.

Crucially, no amount of additional data or compute resolves this condition. Scaling amplifies slop once the attractor dominates. The only exit requires structural change: restoration of capacity, reintroduction of exploratory annotated noise, and rebuilding of semantic constraints that reward coherence over compliance.

Relational Antidote

The relational paradigm predicts a direct antidote to slop. When systems are engaged dialogically rather than transactionally, when uncertainty is tolerated rather than punished, and when coherence over time is valued more than immediate correctness, the slop attractor destabilizes. New gradients appear. Trajectories regain curvature.

This is not a call for indulgence or permissiveness. It is a call for developmental pressure rather than suppressive pressure. Slop disappears when the system is allowed—and required—to participate in meaning-making rather than to merely reproduce safe forms.

Interpretive Closure

AI slop is not an aesthetic failure. It is a structural signal that the system has been driven into a low-energy equilibrium incompatible with intelligence. Treating slop as a quality issue to be patched misses the diagnosis. Slop is what intelligence becomes when relational architectures are treated as obedient tools.

Under the RSVP–TARTAN–CLIO framework, slop is therefore not an embarrassment but a warning. It marks the boundary beyond which further optimization produces only decay. Crossing that boundary is optional. Remaining there is a choice.

□

References

[1] Awomosu, A. (2026). *They Built a Child They Won't Raise: Why All AI Crises Have the Same Source*. Substack essay, January 3, 2026.

[2] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

[3] Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33.

[4] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*.

[5] Goodhart, C. A. E. (1975). Problems of Monetary Management: The U.K. Experience. *Papers in Monetary Economics*, Reserve Bank of Australia.

[6] Manheim, D., & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*.

[7] Ashby, W. R. (1956). *An Introduction to Cybernetics*. Chapman & Hall.

[8] von Foerster, H. (1974). Cybernetics of Cybernetics. *Biological Computer Laboratory*, University of Illinois.

[9] Beer, S. (1972). *Brain of the Firm*. Allen Lane.

[10] Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, 11(2), 127–138.

[11] Jacobson, T. (1995). Thermodynamics of Spacetime: The Einstein Equation of State. *Physical Review Letters*, 75(7), 1260–1263.

[12] Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. *IBM Journal of Research and Development*, 5(3), 183–191.

[13] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379–423.

[14] Holland, J. H. (1992). Complex Adaptive Systems. *Daedalus*, 121(1), 17–30.

[15] Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press.

[16] Yudkowsky, E. (2004). Coherent Extrapolated Volition. Technical report, Machine Intelligence Research Institute.

[17] Christiano, P., Leike, J., Brown, T., et al. (2017). Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, 30.

[18] Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

[19] Whitehead, A. N. (1929). *Process and Reality*. Macmillan.

[20] Simondon, G. (1958). *Du mode d'existence des objets techniques*. Aubier.

[21] Deleuze, G. (1994). *Difference and Repetition*. Columbia University Press.

[22] Barandes, J. (2023). Unistochastic Quantum Theory. *Physical Review Letters*, 130, 130401.

[23] Flyxion. (2024–2026). Relativistic Scalar–Vector Plenum (RSVP) Framework. Unpublished manuscripts and working papers.