Concept Paper

NeuroCore: A Framework for Neuromodulation-Regulated Self-Modifying Modular Neural Architectures

Yijiang Li *

Posted Date: 2 March 2026

doi: 10.20944/preprints202603.0075.v1

Keywords: modular neural architectures; neuromodulation; self-modifying systems; distributional reinforcement learning; meta-reinforcement learning; continual learning; stability theory

 doi:10.20944/preprints202603.0075.v1

Concept Paper

NeuroCore: A Framework for Neuromodulation-Regulated Self-Modifying Modular Neural Architectures

Yijiang Li

Washington University in St.Louis: l.yijiang@wustl.edu

Abstract

We introduce the NeuroCore framework, a formal mathematical treatment of modular neural architectures in which a minimal executive Core—possessing no higher cognitive capabilities—autonomously orchestrates a heterogeneous collection of specialist modules through learned continuous-representation interfaces. The Core's behavior is governed by two neuromodulation-inspired subsystems: a Dopamine System implementing distributional reinforcement learning with prediction-error intrinsic motivation and a stagnation penalty, and a Serotonin System formulated as a meta-reinforcement-learning controller that learns to optimize long-horizon constraint satisfaction. We make four theoretical contributions. First, we formalize the *stagnation-modification tradeoff*—proving that without explicit anti-stagnation pressure, optimal policies in self-modifying systems converge to modification-avoidance, and deriving the conditions under which the stagnation penalty restores non-trivial self-modification behavior (Theorem 1). Second, we prove a *general non-convergence result* for coupled self-modifying multi-objective systems, showing that the joint optimization does not admit guaranteed convergence to fixed points or bounded attractors in the parameter space (Theorem 2). Third, we establish *partial stability guarantees*: bounded representational drift via homeostatic Lyapunov functions (Theorem 3), local convergence under frozen modules via two-timescale stochastic approximation (Proposition 1), and modification frequency bounds (Proposition 2). Fourth, we derive *information-theoretic costs* for module manipulation operations that serve as principled proxies for true disruption. We propose seven falsifiable empirical predictions and discuss implications for the design of autonomous self-organizing AI systems.

Keywords: modular neural architectures; neuromodulation; self-modifying systems; distributional reinforcement learning; meta-reinforcement learning; continual learning; stability theory

1. Introduction

The dominant paradigm in contemporary artificial intelligence—training a monolithic neural network on a static dataset, then deploying it as a frozen inference engine—achieves remarkable task performance but lacks the adaptive, self-organizing properties that characterize biological intelligence. Reinforcement learning (RL) agents exhibit goal-directed autonomy but remain confined to narrow task domains with fixed architectures. AI agent frameworks (AutoGPT, Voyager; Wang et al., 2023) achieve flexible multi-step behavior through large language model (LLM) orchestration, but communicate between components via serialized text, discarding the rich continuous representations that neural networks learn internally.

We argue that progress toward genuinely adaptive AI requires addressing three structural deficits simultaneously: (i) the absence of principled, intrinsically motivated autonomy; (ii) the shallowness of inter-module communication in modular systems; and (iii) the external management of all adaptation and learning. No existing architecture addresses all three.

This paper introduces the NeuroCore framework, a formal mathematical treatment of a modular neural architecture organized around a central thesis: that intelligent self-organization can emerge

 doi:10.20944/preprints202603.0075.v1

from the interaction between a minimal executive controller and a rich ecosystem of specialist modules, provided the controller has appropriate motivational drives (curiosity, anti-stagnation pressure), regulatory capabilities (adaptive constraint satisfaction, homeostatic stabilization), and consequential agency over its own module configuration.

The minimal Core thesis. A defining commitment of NeuroCore is that the Core possesses no higher cognitive capabilities. It cannot understand language, perceive images, reason about physics, or generate content. All such capabilities reside exclusively in specialist modules. The Core provides only autonomy (the capacity and motivation to act), regulation (the capacity to maintain stability and enforce constraints), and action (the capacity to execute decisions in both the external environment and the internal module landscape). This design is motivated by the observation that biological intelligence emerges not from a single monolithic system but from compact subcortical controllers (basal ganglia, brainstem nuclei) orchestrating a diverse ensemble of cortical specialist areas.

Our contributions are as follows:

1. We formalize the *stagnation-modification tradeoff* (§5.3): in systems where self-modification is costly, rational agents converge to modification-avoidance. We prove conditions under which a stagnation penalty restores non-trivial self-modification (Theorem 1).

2. We prove a *general non-convergence theorem* (§7.1) for coupled self-modifying multi-objective systems, establishing that the joint optimization of NeuroCore's subsystems does not admit guaranteed convergence.

3. We establish *partial stability guarantees* (§7.2–7.4): bounded representational drift via a homeostatic Lyapunov function (Theorem 3), local convergence under frozen modules (Proposition 1), and modification frequency bounds (Proposition 2).

4. We derive *information-theoretic module manipulation costs* (§6) that are principled proxies for true representational disruption, and formalize the meta-RL Serotonin System as a generalization of constrained MDPs for non-stationary settings (§5.2).

2. Related Work

Modular neural architectures. Mixture of Experts (MoE; Shazeer et al., 2017; Fedus et al., 2022) routes inputs to specialist sub-networks via a gating function, but all experts share one architecture and are trained jointly. Routing Networks (Rosenbaum et al., 2018) use RL to select among heterogeneous function blocks for multi-task learning. Neural Module Networks (Andreas et al., 2016) compose task-specific architectures from a library of modules. Deep Model Reassembly (Yang et al., 2022) demonstrated that blocks from heterogeneous vision architectures can be stitched together with reasonable performance. NeuroCore differs from all of these in that modules can be dynamically manipulated (fine-tuned, composed, swapped, added, removed) at runtime by an RL-driven controller, and communication occurs through learned continuous-representation interfaces rather than shared token vocabularies.

Neuroscience-inspired AI. Doya (2002) proposed a mapping between neuromodulators and RL meta-parameters: dopamine → reward prediction error, serotonin → temporal discount factor, norepinephrine → inverse temperature (exploration), acetylcholine → learning rate. The Global Workspace Theory (GWT) implementation by VanRullen and Kanai (2021) connects pretrained modules via a shared latent workspace with attention-based routing. Recent work on Lifelong RL via Neuromodulation (2024) demonstrates that neuromodulatory signals can gate plasticity for continual learning. NeuroCore builds directly on Doya's framework but elevates the serotonergic subsystem from a passive parameter to an active meta-RL controller.

Self-modifying systems. Schmidhuber's self-referential weight matrices (1993; Irie et al., 2022) allow networks to modify their own weights at runtime. The Gödel Machine (Schmidhuber, 2007) is a theoretical framework for self-improving systems with provable guarantees. The Darwin Gödel Machine (Sakana AI, 2025) achieves practical self-modifying code. Learning-theoretic analysis (2025) shows that policy-level learnability is preserved only if the policy-reachable family has uniformly

 doi:10.20944/preprints202603.0075.v1

bounded capacity. NeuroCore's self-modification operates at the module level rather than individual weights, and we provide explicit stability analysis for this coarser granularity.

Safe and constrained RL. Constrained MDPs (Altman, 1999) and Constrained Policy Optimization (Achiam et al., 2017) enforce safety constraints via Lagrangian relaxation with reactive multiplier updates. Gao, Schulman, and Hilton (2022) established scaling laws for reward model overoptimization, showing that optimizing against an imperfect proxy inevitably degrades the true objective. Our meta-RL Serotonin System generalizes Lagrangian methods to handle the non-stationarity introduced by self-modification, at the cost of additional optimization complexity.

3. Preliminaries and Notation

We work within the framework of Markov Decision Processes (MDPs) extended to accommodate non-stationarity, multiple objectives, and hierarchical action spaces.

Definition 1 (Non-Stationary Constrained MDP). A non-stationary constrained MDP is a tuple $\mathcal{M}(t) = (S, A, P_t, R_t, \gamma(t), \{C_i\}_{i=1}^m, \{d_i\}_{i=1}^m)$ where S is the state space, A = A_ext ∪ A_int is a hierarchical action space (external and internal actions), $P_t : S \times A \to \Delta(S)$ is a time-varying transition kernel (non-stationary due to self-modification), $R_t : S \times A \to \mathbb{R}$ is a time-varying reward function, $\gamma(t) \in (0, 1)$ is a state-dependent discount factor, $C_i : S \times A \to \mathbb{R}$ are m constraint cost functions, and $d_i \in \mathbb{R}$ are constraint thresholds.

Definition 2 (Distributional Return). The distributional return $Z\pi(s, a)$ is a random variable representing the discounted sum of future rewards under policy π: $Z\pi(s, a) = \Sigma_{t=0}^{\infty} \gamma^t R(s_t, a_t)$ with $s_0 = s$, $a_0 = a$. The distributional Bellman operator $T\pi$ is defined by $T\pi Z(s, a) =^d R(s, a) + \gamma Z(s', a')$ where $=^d$ denotes equality in distribution.

We use the following notation throughout: $h^c \in \mathbb{R}^{d_c}$ for the Core's latent state; π^d for the Dopamine System's policy parameterized by θ^d; π_s for the Serotonin System's meta-policy parameterized by θ_s; τ_i for the T-Interface connecting module M_i to the Core; $\mathcal{E}$ for the module registry; D_KL for Kullback-Leibler divergence; I(X; Y) for mutual information; and CKA for centered kernel alignment.

4. The NeuroCore Framework

Definition 3 (NeuroCore System). A NeuroCore system is a tuple $\Psi = (C, \mathcal{E}, \{\tau_i\}_{i \in \mathcal{E}})$ where $C = (\pi^d, \pi_s, \pi\_B, h^c)$ is the Core (comprising the Dopamine policy, Serotonin meta-policy, Behavior policy, and latent state), $\mathcal{E} = \{M_1, ..., M_n\}$ is the module registry, and $\{\tau_i\}$ are the T-Interfaces. The system state at time t is:

$$\Psi(t) = \langle C(t), \{M_i(t)\}_{i \in \mathcal{E}}, \{\tau_i(t)\}_{i \in \mathcal{E}} \rangle \quad (1)$$

4.1. Module Specification

Definition 4 (Module). A module $M_i = (f_i, d_i^N, d_i^{out}, \theta_i, \omega_i)$ consists of a forward function $f_i : \mathbb{R}^{d_i^N} \to \mathbb{R}^{d_i^{out}}$ (architecture-specific), input and output dimensionalities, learnable parameters θ_i, and metadata ω_i (architecture type, computational cost, task domain). Modules are heterogeneous: the set $\{f_i\}_{i \in \mathcal{E}}$ may include Transformers, CNNs, diffusion models, RNNs, state-space models, graph neural networks, and others. The Core imposes no architectural constraints on modules.

4.2. T-Interfaces

Definition 5 (T-Interface). A T-Interface $\tau_i = (\tau_i^N, \tau_i^{out})$ is a pair of differentiable mappings:

$$\tau_i^N : \mathbb{R}^{d_c} \to \mathbb{R}^{d_i^N} \text{ (Core → Module) (2a)}$$

$$\tau_i^{out} : \mathbb{R}^{d_i^{out}} \to \mathbb{R}^{d_c} \text{ (Module → Core) (2b)}$$

that bridge the Core's representation space and the module's native representation space while preserving gradient flow. The architecture of each τ component depends on the output geometry of the connected module (cross-attention pooling for sequence outputs, spatial attention for spatial

 doi:10.20944/preprints202603.0075.v1

outputs, learned linear projections for vector outputs). T-Interfaces may themselves be parameterized neural networks of arbitrary complexity.

Remark. The essential requirement is that T-Interfaces operate on continuous activations, never on serialized text tokens. If module M_i's native output is text (e.g., an LLM's token sequence), τ_i^{out} operates on the LLM's final hidden-layer activations, not the decoded text. This is what distinguishes NeuroCore from AI agent frameworks: information is transmitted in its native continuous form, not lossy-compressed into discrete symbols.

4.3. Core State Update

At each decision step, the Core aggregates information from active modules via gated attention fusion:

$$z_i(t) = \tau_i^{out}(f_i(x_i(t); \theta_i)) \text{ for each active module } i \in A(t) \subseteq \mathcal{E} \quad (3)$$

$$h^c(t) = \text{CoreRNN}(h^c(t-1), \text{Attn}(h^c(t-1), \{z_i(t)\}_{i \in A(t)})) \quad (4)$$

where CoreRNN is a recurrent state-update function (GRU variant) and Attn computes multi-head attention with the Core state as query and module projections as keys/values. The attention mechanism naturally handles variable numbers of active modules. Note that the Core's 'understanding' of any situation is entirely mediated by modules: it has no direct access to raw sensory input or linguistic content.

5. Neuromodulatory Subsystems

5.1. The Dopamine System

The Dopamine System implements the Core's primary learning and decision-making mechanism through distributional reinforcement learning. Grounded in the empirical finding that biological dopamine neurons encode distributional return predictions (Dabney et al., 2020; Muller et al., 2024), the system maintains a full return distribution via the Implicit Quantile Network (IQN; Dabney et al., 2018):

$$Z\theta(s, a; \tau) = F^{-1}(\tau; s, a), \tau \sim \text{Uniform}(0, 1) \quad (5)$$

trained via the quantile Huber loss:

$$\mathcal{L}\_QR(\theta) = E[\Sigma_i \Sigma_j \varrho^k{}_{\tau_i}(\delta_{ij})], \delta_{ij} = r + \gamma(t)\, Z\theta'(s', a^*; \tau_j) - Z\theta(s, a; \tau_i) \quad (6)$$

Epistemic uncertainty is estimated through a deep ensemble of K distributional critics:

$$U\_epi(s, a) = \text{Var}_k[E_\tau[Z\theta_k(s, a; \tau)]] \quad (7)$$

Intrinsic motivation is derived from world-model prediction error, normalized by the model's own variance estimate:

$$r\_int(t) = \| h^c(t+1) - W\varphi(h^c(t), a(t)) \|^2 / (1 + \sigma^2\varphi(t+1)) \quad (8)$$

The denominator mitigates the 'noisy TV' problem (Burda et al., 2019): inherently stochastic transitions yield both high prediction error and high variance, producing low intrinsic reward. Only systematically incorrect predictions—genuine model ignorance—generate strong motivation.

5.2. The Serotonin System as Meta-RL Controller

The Serotonin System is formulated as a meta-reinforcement-learning controller that learns an adaptive regulatory policy. This design choice is motivated by neuroscience: biological serotonin does not merely enforce fixed thresholds but actively modulates temporal discounting (Doya, 2002), risk sensitivity, learning rates, and the balance between habitual and goal-directed control (Daw et al., 2002)—functions that collectively constitute a learned regulatory policy on slower timescales than dopaminergic processing. Experimentally, Cardozo Pinto et al. (2024) confirmed opponent dopamine-serotonin dynamics using dual-color optogenetics in mouse nucleus accumbens.

Definition 6 (Serotonin Meta-Policy). The Serotonin System's meta-policy π_s parameterized by θ_s outputs a regulatory action vector:

$$a_s(t) = \pi_s(h^c(t), H_t; \theta_s) = (\gamma(t), \alpha(t), \beta(t), \eta\_sc(t), \lambda_1(t), ..., \lambda_m(t)) \quad (9)$$

 doi:10.20944/preprints202603.0075.v1

where H_t is a history buffer of recent Core states and constraint violations. The components are: $\gamma(t) \in (0, 1)$, the temporal discount factor; $\alpha(t)$, $\beta(t) \geq 0$, exploration-exploitation coefficients for prediction-error and epistemic-uncertainty bonuses respectively; $\eta\_sc(t) \in (0, 1]$, the gradient scaling factor controlling Core-to-module gradient flow; and $\lambda_i(t) \geq 0$, adaptive constraint multipliers.

The meta-RL objective minimizes long-horizon aggregate constraint violation:

$$\theta_s \leftarrow \text{argmin}\_\theta_s \, E\_\pi_s[\Sigma_t \, \gamma_s^t \, L\_c(t)] \quad (10)$$

$$L\_c(t) = \Sigma_i \, w_i \, \max(0, C_i(t) - d_i)^2 + \lambda\_d \, D\_KL(\pi^d(t) \parallel \pi\_safe) + \lambda\_h \, \mathcal{L}\_homeo(t) \quad (11)$$

where $\mathcal{L}\_homeo$ is the homeostatic activation regularizer:

$$\mathcal{L}\_homeo = c_1(\mu(h^c) - \mu_0)^2 + c_2(\sigma(h^c) - \sigma_0)^2 + c_3 \, D\_KL(p(h^c) \parallel p_0) \quad (12)$$

with target statistics μ_0, σ_0, p_0 established during initial training (analogous to biological homeostatic set points; Turrigiano, 2008).

Advantage over standard constrained MDPs. Lagrangian methods update constraint multipliers reactively: increasing λ after violations, decreasing after compliance. In NeuroCore, where module modifications introduce discontinuous changes in the transition dynamics, reactive methods must re-discover appropriate multipliers after each modification, producing oscillatory constraint violations. The meta-RL formulation learns a policy that can anticipate constraint violations from behavioral precursors in h^c and preemptively adjust regulation. The cost is a second RL optimization loop with its own convergence challenges (§7).

5.3. The Stagnation-Modification Tradeoff

We now formalize a fundamental problem in self-modifying systems: if self-modification is costly and potentially destructive, rational agents will converge to never modifying.

Definition 7 (Modification-Avoidant Policy). A policy π is modification-avoidant if $\Pr(a \in A\_int \mid \pi) < \varepsilon$ for small $\varepsilon > 0$, i.e., it essentially never selects internal (self-modification) actions.

Theorem 1 (Stagnation Convergence). Let $\mathcal{M}$ be a non-stationary constrained MDP with action space $A = A\_ext \cup A\_int$, where each internal action $a \in A\_int$ incurs immediate cost $C(a) > 0$ and causes a stochastic perturbation to the transition dynamics P_t. If (i) the expected post-modification return $E[V\pi'(s)]$ is uncertain with variance $\sigma^2\_mod > 0$, (ii) the discount factor satisfies $\gamma < 1$, and (iii) there exists a viable external-only policy π_ext with $V\pi^{ext}(s) > 0$ for some states, then for sufficiently small learning rates, the policy gradient update converges to a modification-avoidant policy π^* with $\Pr(a \in A\_int \mid \pi^*) \to 0$.

Proof. Consider the expected return of an internal action a_m at state s versus the best external action a_e:

$$Q(s, a\_m) = -C(a\_m) + \gamma \, E[V\pi(s' \mid a\_m)] \quad (13)$$

$$Q(s, a\_e) = R(s, a\_e) + \gamma \, E[V\pi(s' \mid a\_e)] \quad (14)$$

The modification action incurs immediate cost $C(a\_m) > 0$ and perturbs the MDP dynamics, meaning the agent's current value function $V\pi$ becomes inaccurate for the post-modification MDP. By the performance difference lemma (Kakade & Langford, 2002), the return under the current policy in the perturbed MDP satisfies:

$$|V\pi\_\mathcal{M}' - V\pi\_\mathcal{M}| \leq (2\gamma / (1-\gamma)) \cdot \max\_s \, |E\_\mathcal{M}'[R|s] - E\_\mathcal{M}[R|s]| \quad (15)$$

Since the perturbation's direction and magnitude are uncertain, the agent's estimate of $Q(s, a\_m)$ has variance proportional to $\sigma^2\_mod / (1 - \gamma)^2$, which is amplified by the discount factor for long horizons. Under policy gradient methods, actions with high-variance returns and guaranteed immediate costs have their gradients dominated by the cost term. Formally, the gradient contribution is:

$$\nabla\_\theta \log \pi(a\_m|s) \cdot (Q(s, a\_m) - b(s)) \approx \nabla\_\theta \log \pi(a\_m|s) \cdot (-C(a\_m) + \text{noise}) \quad (16)$$

The expected gradient is negative (pushing probability mass away from a_m), while the noise has zero mean. Over sufficient iterations, $\pi(a\_m \mid s) \to 0$ for all modification actions, yielding a modification-avoidant policy. ■

 doi:10.20944/preprints202603.0075.v1

Corollary 1. The modification-avoidant equilibrium is locally stable: small perturbations that occasionally trigger modifications will be corrected by the policy gradient, returning the system to avoidance.

To break the stagnation equilibrium, we introduce an anti-stagnation mechanism based on the rate of epistemic uncertainty reduction. Let $U(t) = E_s[U_{epi}(s, a)]$ be the average epistemic uncertainty:

$$r_{stag}(t) = -\zeta \cdot \max(0, U(t) - U(t - \Delta T) + \varepsilon_{min}) \quad (17)$$

Proposition (Stagnation penalty sufficiency). If $\zeta > C_{max} / (\varepsilon_{min} \cdot \Delta T)$, where $C_{max} = \max_a C(a)$, then the modification-avoidant policy is no longer a local optimum of the augmented objective $r_{total} = r_{ext} + \alpha\, r_{int} + \beta\, U_{epi} + r_{stag}$.

Proof sketch. Under a modification-avoidant policy, epistemic uncertainty in unvisited regions does not decrease (the system exploits only known states). After ΔT steps without uncertainty reduction, the stagnation penalty contributes $-\zeta \cdot \varepsilon_{min}$ per step. The cumulative penalty over a horizon of H steps is $-\zeta \cdot \varepsilon_{min} \cdot H$. For $\zeta > C_{max} / (\varepsilon_{min} \cdot \Delta T)$, a single modification action at cost C_{max} followed by uncertainty reduction exceeding ε_{min} produces higher expected return than continued stagnation for $H > C_{max} / (\zeta \cdot \varepsilon_{min})$, creating a policy gradient toward non-trivial modification. ■

We additionally introduce a metabolic utilization cost for dormant modules:

$$r_{util}(t) = -\zeta_2 \cdot H(\text{dormant}(t)) \quad (18)$$

where H(dormant) measures the information-theoretic cost of maintaining modules that have not contributed to the Core state in the past N steps, creating pressure to either utilize or prune unused modules.

The composite reward signal is:

$$r_{total}(t) = r_{ext}(t) + \alpha(t)\, r_{int}(t) + \beta(t)\, U_{epi}(t) + r_{stag}(t) + r_{util}(t) \quad (19)$$

where $\alpha(t)$ and $\beta(t)$ are regulatory outputs of the Serotonin meta-policy (Eq. 9).

5.4. Dual-Timescale Dynamics

The Dopamine and Serotonin systems operate on separated timescales:

$$\theta^d(t+1) = \theta^d(t) - \eta^d \nabla \mathcal{L}^d(\theta^d; a_s(t)) \quad (20a)$$

$$\theta_s(t+K) = \theta_s(t) - \eta_s \nabla \mathcal{L}_{meta}(\theta_s),\ \eta_s \ll \eta^d,\ K \in [10, 50] \quad (20b)$$

This separation is both biologically motivated (serotonergic modulation operates on seconds-to-minutes timescales versus millisecond dopaminergic firing) and mathematically convenient, enabling analysis via two-timescale stochastic approximation theory (Borkar, 2008).

6. Information-Theoretic Self-Modification Costs

A principled cost function for module manipulation must reflect the true representational disruption caused by the manipulation. We derive such costs from information-theoretic quantities.

Definition 8 (Integration Depth). The integration depth of module M_i with the Core is measured by the mutual information between the module's T-Interface output and the Core's state update:

$$D_i = I(z_i(t); h^c(t) \mid h^c(t-1), \{z_j(t)\}_{j \neq i}) \quad (21)$$

This measures how much information module M_i uniquely contributes to the Core state, conditional on the Core's prior state and all other modules. A module with $D_i \approx 0$ is redundant; a module with high D_i is irreplaceable.

Module manipulation costs are derived from D_i:

$$C(\text{remove}, M_i) = \kappa_1 \cdot D_i \quad (22a)$$

$$C(\text{swap}, M_i, M_i') = \kappa_2 \cdot (D_i + D_{CKA}(M_i, M_i')) \quad (22b)$$

$$C(\text{finetune}, M_i, n) = \kappa_3 \cdot n \cdot D_i^{1/2} \quad (22c)$$

$$C(\text{adjust}_\tau, \tau_i, \Delta\theta) = \kappa_4 \cdot \|\Delta\theta\|^2 \quad (22d)$$

where D_{CKA} is the centered kernel alignment distance between the two modules' representation spaces, n is the number of fine-tuning steps, and $\kappa_1 \ldots \kappa_4$ are scaling constants. The

 doi:10.20944/preprints202603.0075.v1

square-root dependence in Eq. 22c reflects the empirical observation that fine-tuning disruption grows sublinearly with the number of steps (due to LoRA's low-rank constraint).

Proposition (Cost-disruption correlation). Under mild regularity conditions (Lipschitz continuity of f_i and τ_i), the removal cost C(remove, M_i) is a lower bound on the expected performance degradation:

$$E[\Delta V \mid \text{remove } M_i] \geq \alpha_0 \cdot D_i - \varepsilon\_approx \quad (23)$$

where α_0 depends on the Core's reliance on module i's information and ε_approx is an approximation error from finite-sample mutual information estimation. This justifies D_i as a meaningful proxy for true disruption.

7. Stability Analysis

This section provides the paper's central theoretical results on the stability and convergence properties of the NeuroCore system. We first establish a negative result (non-convergence in general), then derive three partial stability guarantees that collectively characterize the system's operating regime.

7.1. Non-Convergence Theorem

Theorem 2 (General non-convergence). Let $\Psi(t)$ be the NeuroCore system state. The joint optimization of $(\pi^d, \pi_s, \{\tau_i\}, \{\theta_i\})$ does not, in general, converge to a fixed point, limit cycle, or bounded attractor in the parameter space.

Proof. We establish non-convergence via three independent mechanisms, any one of which suffices.

(a) Self-modification-induced non-stationarity. Let $T_1, T_2, \ldots$ be the sequence of module modification events. At each T_k, the Core modifies some module M_i, transitioning the system from MDP $\mathcal{M}_k = (S, A, P_k, R_k, \gamma_k)$ to $\mathcal{M}_{k+1}$. The value function $V\pi^d$ optimized under $\mathcal{M}_k$ is generally invalid under $\mathcal{M}_{k+1}$. By the simulation lemma (Kearns & Singh, 2002), the value function error satisfies:

$$|V\pi\_\{\mathcal{M}_k\} - V\pi\_\{\mathcal{M}_{k+1}\}| \leq (2\gamma\_max\ R\_max) / (1 - \gamma\_max)^2 \cdot max\_s\ D\_TV(P_k(\cdot|s,a), P_{k+1}(\cdot|s,a)) \quad (24)$$

where D_TV is the total variation distance. Unless the modification process itself converges (i.e., the agent eventually stops modifying), the MDP sequence $\{\mathcal{M}_k\}$ does not converge, and neither does the value function. But proving convergence of the modification process requires proving convergence of the overall system—a circularity.

(b) Two-player non-cooperative game. The Dopamine System maximizes r_total (which includes returns from potentially constraint-violating actions), while the Serotonin System minimizes constraint violations (which requires suppressing some high-return behaviors). This constitutes a two-player general-sum game. Daskalakis, Goldberg, and Papadimitriou (2009) proved that computing Nash equilibria in general-sum games is PPAD-complete. Mazumdar, Ratliff, and Sastry (2020) showed that gradient dynamics in continuous games do not converge to Nash equilibria in general, and can exhibit limit cycles or chaotic trajectories in parameter space even for simple bilinear games. The NeuroCore game is substantially more complex (nonlinear, high-dimensional, non-stationary), precluding convergence guarantees.

(c) Non-convex individual objectives. Each subsystem's optimization involves deep neural networks. Even for single-agent deep RL with nonlinear function approximation, Tsitsiklis and Van Roy (1997) demonstrated divergence of temporal-difference learning. The coupling between multiple such systems—where each system's loss landscape depends on the other systems' parameters—creates a composite landscape where saddle points, local minima, and degenerate plateaus are ubiquitous. ■

 doi:10.20944/preprints202603.0075.v1

7.2. Bounded Representational Drift

Theorem 3 (Homeostatic drift bound). Let $p(h^c(t))$ denote the distribution of Core activations at time t and p_0 the initial distribution. If the homeostatic regularization coefficient c_3 in Eq. 12 satisfies $c_3 \geq c^*(\varepsilon, G\_max)$, where $G\_max = \sup_t \|\nabla_{h^c} \mathcal{L}^d\|$ is the maximum policy gradient magnitude, then:

$$D_{KL}(p(h^c(t)) \| p_0) \leq \varepsilon \text{ for all } t \quad (25)$$

Proof. Define the Lyapunov function $V(t) = D_{KL}(p(h^c(t)) \| p_0)$. The Core state update (Eq. 4) at each step applies a parameterized transformation to h^c, causing a distributional shift. By Pinsker's inequality and the data processing inequality, the KL-shift per step is bounded:

$$\Delta V(t) \leq \eta^d \cdot G\_max \cdot \sqrt{(2V(t))} \quad (26)$$

The homeostatic loss $\mathcal{L}\_homeo$ contributes a restoring gradient:

$$\Delta V\_homeo(t) \leq -\eta^d \cdot c_3 \cdot V(t) \quad (27)$$

Combining, the net drift satisfies:

$$\Delta V(t) \leq \eta^d \left(G\_max \sqrt{(2V(t))} - c_3 V(t)\right) \quad (28)$$

Setting $\Delta V(t) = 0$ and solving for the equilibrium gives $V^* = 2G^2\_max / c^2_3$. For $V^* \leq \varepsilon$, we require $c_3 \geq G\_max \sqrt{(2/\varepsilon)} =: c^*(\varepsilon, G\_max)$. When $V(t) > V^*$, the restoring term dominates the drift term, pulling V(t) back below the bound. ■

Remark. Theorem 3 provides a tunable tradeoff: stronger homeostatic regulation (larger c_3) yields tighter drift bounds but constrains the Core's representational plasticity. The Serotonin meta-policy can learn to balance this tradeoff by adjusting the effective strength of homeostatic regulation across different phases of operation.

7.3. Local Convergence Under Frozen Modules

Proposition 1 (Two-timescale local convergence). If the module configuration is held fixed ($\mathcal{E}$ constant, no internal actions), the NeuroCore system reduces to a two-timescale stochastic approximation:

$$\theta^d(t+1) = \theta^d(t) + \eta^d \left(g^d(\theta^d, \theta_s) + M^d(t+1)\right) \quad (29a)$$

$$\theta_s(t+1) = \theta_s(t) + \eta_s \left(g_s(\theta^d, \theta_s) + M_s(t+1)\right) \quad (29b)$$

where g^d, g_s are the expected gradient mappings and M^d, M_s are martingale difference noise terms. Under standard conditions—(i) Lipschitz continuity of g^d and g_s, (ii) $\eta^d / \eta_s \to \infty$, (iii) standard step-size conditions $\Sigma\eta = \infty$ and $\Sigma\eta^2 < \infty$, and (iv) bounded iterates—the fast system θ^d tracks the quasi-static equilibrium $\theta^{*d}(\theta_s)$ and the slow system θ_s evolves on the equilibrium manifold (Borkar, 2008, Theorem 2, Chapter 6). Local convergence to a locally asymptotically stable equilibrium of the slow-timescale ODE is guaranteed.

Proof sketch. Under frozen modules, the MDP is stationary: $P_t = P$, $R_t = R$. The fast system (Dopamine) sees a fixed MDP with fixed regulatory signals and satisfies standard policy gradient convergence conditions for the tabular/compatible function approximation case. The slow system (Serotonin) observes the time-averaged behavior of the converged fast system and optimizes a smooth objective over it. The two-timescale framework of Borkar (2008) and Bhatnagar et al. (2009) applies directly. ■

7.4. Self-Modification Frequency Bound

Proposition 2 (Modification frequency bound). Let $N\_mod(T) = |\{T_k : T_k \in [0, T]\}|$ be the number of module modifications in the interval [0, T]. Under the constraint budget B and minimum modification cost $C\_min = \min_{a \in A\_int} C(a)$, the expected modification frequency is bounded:

$$E[N\_mod(T)] \leq B \cdot T / C\_min \quad (30)$$

Moreover, the inter-modification interval $\Delta T\_mod = T_{k+1} - T_k$ satisfies $E[\Delta T\_mod] \geq C\_min / B$.

Proof. Each modification action consumes at least C_min from the per-step constraint budget B. The Serotonin System's constraint enforcement (Eq. 10–11) ensures that the long-run average constraint cost is bounded by B. By the renewal reward theorem, the long-run modification rate is at most B / C_min. ■

 doi:10.20944/preprints202603.0075.v1

Combined with Proposition 1, this result establishes that the system spends most of its time in the frozen-module regime where local convergence applies, punctuated by infrequent modification events that perturb the system to a new operating point. The key design question is whether the inter-modification intervals E[ΔT_mod] are long enough for the fast system to approximately reconverge—a condition that depends on the relative magnitudes of C_min, B, and the fast system's convergence rate.

7.5. Summary of Stability Properties

The stability landscape of NeuroCore can be summarized as follows. Globally, the system does not converge (Theorem 2). Locally, during inter-modification intervals with frozen modules, the two-timescale system converges to local equilibria (Proposition 1). The Core's activation distribution remains within a bounded KL-ball around its initial distribution (Theorem 3). The modification frequency is bounded (Proposition 2), ensuring that non-stationarity is controlled. The system is best characterized not as convergent but as *metastable*: it occupies a sequence of quasi-stable operating points, transitioning between them via infrequent self-modification events, with homeostatic regulation preventing catastrophic drift during transitions.

8. Testable Predictions

We formulate seven falsifiable predictions, each specifying a predicted outcome, comparison condition, and metric. Hypotheses are categorized as core (failure undermines the framework) or auxiliary (failure indicates a specific design revision).

H1 (Core). Continuous-representation T-Interfaces yield ≥15% higher cumulative reward and ≥2× faster learning on multi-modal tasks compared to text-serialized communication, controlling for total parameter count.

H2 (Core). After 10^6 environment steps, the Core's modification policy will exhibit entropy $H \geq 1.5$ nats and long-run return exceeding a frozen-module baseline by ≥10%, demonstrating non-trivial emergent self-modification.

H3 (Core). The meta-RL Serotonin System produces ≥30% fewer cumulative constraint violations than a Lagrangian baseline under non-stationarity induced by periodic forced module modifications.

H4 (Auxiliary). Removing the stagnation penalty ($\zeta = 0$) causes internal action frequency to fall below 10^{-4} per step, confirming Theorem 1's prediction of modification avoidance.

H5 (Auxiliary). The integration depth D_i (Eq. 21) correlates with actual performance degradation upon module removal with $r \geq 0.7$, validating the information-theoretic cost function.

H6 (Auxiliary). The KL-divergence bound from Theorem 3 holds empirically for ≥95% of training, with excursions limited to brief transients following module modifications.

H7 (Core). NeuroCore with a minimal Core orchestrating specialist modules outperforms a monolithic model of equivalent parameter count on a heterogeneous task suite, with ≥40% lower inter-task performance variance.

9. Discussion

On the minimal Core thesis. One might object that a Core without higher cognitive capabilities cannot make intelligent self-modification decisions. We counter that the Core does not need to 'understand' its modules cognitively—it learns through reinforcement which abstract actions in which abstract states lead to improved long-run return. The analogy is to the basal ganglia, which orchestrate cortical function without performing cortical computation. The intelligence is in the learned policy, not in explicit comprehension.

On meta-RL versus constrained MDPs. The meta-RL Serotonin System is the framework's most speculative element. It is more expressive than Lagrangian methods and better aligned with biological serotonin's known functions, but introduces a second optimization loop with its own

 doi:10.20944/preprints202603.0075.v1

convergence challenges. If H3 fails empirically, a hybrid design—meta-RL for temporally extended regulatory decisions, Lagrangian methods for hard safety constraints—represents a principled fallback.

On non-convergence. Theorem 2 is not a flaw but a fundamental property of any simultaneously self-modifying and multi-objective system. Biological brains also do not converge to fixed states; they exhibit lifelong plasticity and perpetual non-stationarity. The partial stability results (Theorem 3, Propositions 1–2) collectively establish that the system operates in a bounded, metastable regime—not convergent, but constrained. Whether this metastable regime supports productive behavior is an empirical question addressable through our testable predictions.

Limitations. The framework has several limitations. First, the homeostatic bound (Theorem 3) assumes bounded policy gradients, which may not hold during training instabilities. Second, the mutual information cost (Eq. 21) requires estimation from finite samples, introducing approximation error. Third, the meta-RL Serotonin System's convergence is not guaranteed even in the frozen-module regime (it may cycle). Fourth, and most fundamentally, the 'emergence' hypothesis—that effective self-modification strategies will arise from the interplay of stagnation pressure and modification costs—remains entirely unvalidated. The framework provides the substrate; whether emergence occurs is unknown.

Safety implications. A self-modifying system with autonomous agency raises significant safety concerns. The Serotonin System provides learned regulation, the homeostatic bounds limit drift, and the modification costs create inertia. However, the fundamental alignment challenge—ensuring a self-modifying system's objectives remain aligned with human values as the system changes itself—is not solved by this framework. We note that the architecture's modularity provides natural intervention points: human operators can freeze modules, adjust constraint budgets, or disable self-modification without disrupting the system's ability to operate as a fixed architecture.

10. Conclusion

We have presented the NeuroCore framework: a formal treatment of modular neural architectures in which a minimal executive Core, equipped with distributional RL, a meta-RL regulatory system, and consequential self-modification capabilities, orchestrates a heterogeneous collection of specialist modules through continuous-representation interfaces. Our theoretical contributions establish the stagnation-modification tradeoff (Theorem 1), the fundamental non-convergence of coupled self-modifying systems (Theorem 2), and partial stability guarantees that collectively characterize a bounded metastable operating regime (Theorem 3, Propositions 1–2). The information-theoretic cost functions provide principled proxies for self-modification disruption.

The framework's central thesis—that intelligent self-organization emerges from the interplay between a minimal controller and a rich modular ecosystem, given appropriate motivational and regulatory mechanisms—is a testable claim, not an established fact. We have provided seven falsifiable predictions to support or refute this thesis. The mathematics establishes what is guaranteed (bounded drift, bounded modification frequency, local convergence between modifications) and what is not (global convergence, emergence of effective strategies). This honest delineation of theoretical boundaries is, we believe, as valuable as the positive results.

Appendix A: Notation Summary

Symbol	Description
$h^c \in \mathbb{R}^{d_c}$	Core latent state vector
π^d, θ^d	Dopamine System policy and parameters
π_s, θ_s	Serotonin System meta-policy and parameters
$M_i = (f_i, d_i^N, d_i^{out}, \theta_i, \omega_i)$	Module specification tuple

 doi:10.20944/preprints202603.0075.v1

$\tau_i = (\tau_i^N, \tau_i^{out})$	T-Interface (bidirectional learned transformation)
$\mathcal{E}$	Module registry
A = A_ext ∪ A_int	Hierarchical action space
Zθ(s, a; τ)	Distributional return (IQN quantile function)
U_epi(s, a)	Epistemic uncertainty (ensemble variance)
$D_i = I(z_i; h^e \mid ...)$	Integration depth (conditional mutual information)
γ(t), α(t), β(t), η_sc(t)	Serotonin regulatory outputs
$\lambda_i(t)$	Adaptive constraint multipliers
$\mathcal{L}$_homeo	Homeostatic activation regularization loss
r_stag, r_util	Stagnation and utilization penalties
D_KL, D_TV, D_CKA	KL divergence, total variation, centered kernel alignment

References

Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained Policy Optimization. In Proc. ICML.

Altman, E. (1999). Constrained Markov Decision Processes. Chapman & Hall/CRC.

Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural Module Networks. In Proc. CVPR.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., & Lee, M. (2009). Natural Actor-Critic Algorithms. Automatica, 45(11), 2471–2482.

Borkar, V. S. (2008). Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press.

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2019). Exploration by Random Network Distillation. In Proc. ICLR.

Cardozo Pinto, D. F., et al. (2024). Dopamine and serotonin oppositely modulate reinforcement learning. Nature.

Dabney, W., Rowland, M., Bellemare, M. G., & Munos, R. (2018). Distributional RL with Quantile Regression. In Proc. AAAI.

Dabney, W., et al. (2020). A distributional code for value in dopamine-based RL. Nature, 577, 671–675.

Daskalakis, C., Goldberg, P. W., & Papadimitriou, C. H. (2009). The Complexity of Computing a Nash Equilibrium. SIAM J. Comp., 39(1), 195–259.

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. Neural Networks, 15(4–6), 603–616.

Doya, K. (2002). Metalearning and neuromodulation. Neural Networks, 15(4–6), 495–506.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch Transformers: Scaling to Trillion Parameter Models. JMLR, 23(120), 1–39.

Gao, L., Schulman, J., & Hilton, J. (2022). Scaling Laws for Reward Model Overoptimization. In Proc. ICML.

Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. In Proc. ICLR.

Irie, K., Schlag, I., Csordás, R., & Schmidhuber, J. (2022). A Modern Self-Referential Weight Matrix. In Proc. ICML.

Kakade, S. & Langford, J. (2002). Approximately Optimal Approximate Reinforcement Learning. In Proc. ICML.

Kearns, M. J. & Singh, S. P. (2002). Near-Optimal Reinforcement Learning in Polynomial Time. Machine Learning, 49(2), 209–232.

Mazumdar, E., Ratliff, L. J., & Sastry, S. S. (2020). On Gradient-Based Learning in Continuous Games. SIAM J. Math. of Data Science, 2(1), 103–131.

Muller, T. H., et al. (2024). Distributional reinforcement learning in prefrontal cortex. Nature Neuroscience.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven Exploration by Self-Supervised Prediction. In Proc. ICML.

Rosenbaum, C., Klinger, T., & Riemer, M. (2018). Routing Networks. In Proc. ICLR.

Schmidhuber, J. (1993). A 'Self-Referential' Weight Matrix. In Proc. ICANN, 446–450.

 doi:10.20944/preprints202603.0075.v1

Schmidhuber, J. (2007). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In Artificial General Intelligence, 199–226. Springer.

Shazeer, N., et al. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In Proc. ICLR.

Tsitsiklis, J. N. & Van Roy, B. (1997). An Analysis of Temporal-Difference Learning with Function Approximation. IEEE Trans. Automatic Control, 42(5), 674–690.

Turrigiano, G. G. (2008). The Self-Tuning Neuron: Synaptic Scaling of Excitatory Synapses. Cell, 135(3), 422–435.

VanRullen, R. & Kanai, R. (2021). Deep Learning and the Global Workspace Theory. Trends in Neurosciences, 44(9), 692–704.

Wang, G., et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv:2305.16291.

Yang, J., et al. (2022). Deep Model Reassembly. In Proc. NeurIPS.