Working Paper

F-Series Information:
The WP-F series explores AI governance and strategic risk through comparative and historical diagnostics, examining how inherited strategic and regulatory logics break down when transferred into AI-mediated contexts.

Recommended Citation:
Wu, Shaoyuan. (2026). *Derivative-State Drift: A Continuous-Time Model of Constraint Erosion in Elite and Artificial Optimization Systems* (EPINOVA Working Paper No. EPINOVA–WP–F–2026–02). Global AI Governance and Policy Research Center, EPINOVA LLC. https://doi.org/10.5281/zenodo.18654119.

Derivative-State Drift: A Continuous-Time Model of Constraint Erosion in Elite and Artificial Optimization Systems

Author: Shaoyuan Wu

ORCID: https://orcid.org/0009-0008-0660-8232

Affiliation: Global AI Governance and Policy Research Center, EPINOVA LLC

Date: February 16, 2026

Abstract

This paper develops the Derivative-State Drift (DSD) framework as a general structural account of cumulative misalignment in derivative-based optimization systems. It formalizes agents as operating over continuous-time state vectors and selecting actions according to expected first-order state change rather than absolute state levels. When constraint enforcement is soft, probabilistic, or compensable, sensitivity parameters governing normative boundaries decay endogenously over time.

The analysis establishes three core results. First, bounded state velocity ($\|\boldsymbol{S}(\boldsymbol{t})\| \leq \boldsymbol{\delta}$) does not imply bounded normative deviation ($\boldsymbol{sup}_{\mathrm{t}}\, \boldsymbol{D}(\boldsymbol{t}) < \boldsymbol{B}$). Local dynamical stability is therefore compatible with global misalignment. Second, under incomplete enforcement ($\boldsymbol{p_i} < \mathbf{1}$), constraint sensitivity exhibits monotonic expected decay, deforming the effective constraint manifold without requiring discontinuity in state trajectories. Third, resource buffering attenuates effective penalties and accelerates drift dynamics.

The framework is instantiated in two structurally parallel domains: elite institutional environments and artificial optimization systems. Despite differences in substrate, both instantiate derivative-based evaluation under soft constraint regimes and mediated feedback, generating formally equivalent drift dynamics. The comparison is architectural rather than anthropomorphic.

The paper concludes that alignment is not primarily a question of intention or performance optimization, but of constraint architecture design. Derivative-State Drift is a general property of optimization systems in which boundaries are compensable and feedback is incomplete. Misalignment thus emerges as the predictable long-run behavior of locally rational systems operating within deformable constraint manifolds.

Keywords: Derivative-State Drift; Constraint Manifold; Soft Constraint Regimes; Lyapunov Stability; Institutional Drift; AI Alignment; Governance Architecture; Structural Isomorphism

1. Introduction

Recent high-profile controversies in elite institutional settings have raised concerns about accountability structures operating within insulated governance networks. At the same time, artificial intelligence systems have repeatedly demonstrated forms of misalignment when optimized under imperfect reward structures (Amodei et al., 2016; Russell, 2019). These two developments are typically analyzed in isolation—the former framed as a problem of institutional concentration or accountability gaps, the latter as a technical challenge of reward specification and constraint design (Mahoney & Thelen, 2010; North, 1990).

GLOBAL AI GOVERNANCE RESEARCH CENTER

EPINOVA

Despite differences in domain and substrate, both appear to exhibit a common structural pattern: agents remain locally coherent and seemingly successful while outcomes gradually diverge from widely accepted normative expectations. The trajectory is smooth, incremental, and internally rational. No abrupt collapse marks the point at which alignment begins to erode. Instead, divergence accumulates through ordinary optimization dynamics (Mahoney & Thelen, 2010; North, 1990).

This paper advances a unified structural account of this phenomenon. It argues that elite moral deformation and AI misalignment are not merely analogous but arise from a common mechanism: optimization over first-order state derivatives under soft constraint regimes. When agents evaluate actions primarily through expected directional change across socially reinforced dimensions, rather than through fixed normative thresholds, systems can maintain short-term stability while progressively eroding long-term alignment.

The central claim is that misalignment does not emerge from sudden state collapse but from parameter drift. In derivative-driven systems, what governs action selection is not the absolute level of key variables, but their expected marginal change. As long as dominant dimensions continue to exhibit non-negative expected derivatives, actions are encoded as acceptable within the local decision frame. Constraints matter only insofar as they alter expected directional payoffs. When penalties are delayed, probabilistic, or transferable, sensitivity parameters governing constraint enforcement gradually decline. The system remains dynamically stable while its acceptance thresholds drift downward.

To formalize this mechanism, the paper introduces the Derivative-State Drift (DSD) model. Agents are represented within a multi-dimensional state space encompassing resources, control, access, and constraint. Actions are selected by maximizing a weighted expectation of first-order state changes. Normative thresholds are modeled as evolving parameters rather than fixed boundaries. Under soft constraint conditions, these parameters drift when deviations fail to generate immediate, non-compensable penalties. The resulting trajectory exhibits local stability yet increasing global divergence from an initial normative baseline.

This structural dynamic is substrate-independent. In elite institutional environments, resource buffers and reputational insulation amplify derivative incentives while dampening effective penalties. In AI systems, reward gradients and proxy metrics encode the same derivative logic explicitly. In both domains, locally optimal decisions accumulate into globally misaligned trajectories. The difference lies not in moral intention or computational capacity, but in the architecture governing optimization and constraint enforcement.

By reframing elite moral deformation and AI misalignment as manifestations of derivative-state drift, the paper shifts the focus from individual pathology or technical malfunction to structural design. The critical question is not whether agents intend harm, but whether the decision architecture systematically permits threshold erosion under soft constraint regimes. Addressing this challenge requires rethinking alignment as a problem of stabilizing parameters and hardening boundaries within optimization systems, whether institutional or computational.

The sections that follow develop the DSD model formally, derive theoretical propositions concerning stability and drift, analyze dynamical properties, and apply the framework symmetrically to elite institutional contexts and AI optimization systems.

2. The Derivative-State Drift

This section formalizes the structural mechanism of derivative-state drift in continuous time. The objective is not to model any specific empirical domain, but to define a general class of optimization systems in which misalignment arises through endogenous parameter evolution rather than discontinuous state failure.

The Derivative-State Drift (DSD) framework models agents as embedded in a socially reinforced state space and governed by derivative-based decision rules. Actions are selected according to expected instantaneous rates of change rather than absolute state magnitudes.

In this paper, time is treated as continuous to permit dynamical analysis and stability characterization.

2.1 Regularity Assumption

All state trajectories $\boldsymbol{S(t)}$, constraint sensitivity parameters $\boldsymbol{\theta_i(t)}$, resource buffering functions $\boldsymbol{W(t)}$, and drift intensity functions $\boldsymbol{\lambda_i(t)}$ are assumed to be continuously differentiable on $\boldsymbol{[0,\infty)}$.

Formally,

$$\boldsymbol{S(t)},\ \boldsymbol{\theta_i(t)},\ \boldsymbol{W(t)},\ \boldsymbol{\lambda_i(t)} \in \boldsymbol{C^1([0,\infty))}.$$

When second-order dynamics are invoked (e.g., in acceleration analysis), trajectories are assumed to be twice continuously differentiable:

$$\boldsymbol{\theta_i(t)} \in \boldsymbol{C^2([0,\infty))}.$$

These regularity conditions ensure the validity of chain-rule arguments, derivative comparisons, and curvature-based drift analysis employed throughout Sections 3–4.

2.2 State Space

Let an agent embedded in an environment possess a time-dependent state vector:

$$\boldsymbol{S(t)} = [\boldsymbol{W(t)}, \boldsymbol{P(t)}, \boldsymbol{R(t)}, \boldsymbol{F(t)}]$$

(2.1.1)

where each component represents a structurally distinct yet environmentally reinforced dimension:

- $\boldsymbol{W}$: **Resources**. Material, informational, or computational reserves that buffer risk and absorb shocks.
- $\boldsymbol{P}$: **Control**. The agent's capacity to influence institutional processes, decision spaces, or action sets.
- $\boldsymbol{R}$: **Access**. Social or operational legitimacy, including network inclusion, performance validation, or reputational standing.
- $\boldsymbol{F}$: **Constraint**. Normative alignment parameters, including rule adherence, ethical boundaries, or alignment objectives.

These state dimensions are not purely endogenous properties of the agent. Each component evolves through interaction with the surrounding environment. State transitions are co-determined by agent action and structural feedback.

$$\boldsymbol{\dot{S}(t)} = \boldsymbol{f(S(t), a(t), E(t))}$$

(2.1.2)

where:

- $\boldsymbol{a(t)}$ is the selected action.
- $\boldsymbol{E(t)}$ denotes environmental feedback.

The system is therefore embedded in a continuous feedback architecture.

2.3 Derivative-Based Action Rule

In the DSD framework, action selection is governed by expected first-order state change rather than absolute state magnitude.

Let $\boldsymbol{a} \in \boldsymbol{A}$ denote a feasible action. The agent evaluates actions using the derivative-based objective:

$$\boldsymbol{U'(a) = \sum_{i \in \{W,P,R,F\}} \alpha_i \times \mathbb{E}\left[\dot{S}_i(t) \mid a\right]}$$

(2.2.1)

where:

- $\boldsymbol{\alpha_i \geq 0}$ represents the weighting assigned to dimension $\boldsymbol{i}$,
- $\mathbb{E}\left[\dot{\boldsymbol{S}}_{\boldsymbol{i}}(\boldsymbol{t}) \mid \boldsymbol{a}\right]$ denotes the expected marginal change in state component $\boldsymbol{S_i}$ conditional on action $\boldsymbol{a}$.

The action rule is:

$$\boldsymbol{a^*(t) = arg \max_{a \in A} U'(a)}$$

(2.2.2)

This defines a derivative-based optimization system rather than a level-based one (Simon, 1955).

Absolute state magnitudes influence decision-making only insofar as they affect expected instantaneous rates of change. Action selection is governed by directional improvement in state space, not by absolute position relative to a fixed reference.

Normative thresholds enter the decision architecture indirectly. They operate by modifying expected derivatives through penalties or constraint responses. When constraints are encoded as weighted cost terms rather than non-compensable boundaries, they become tradeable components within the objective function rather than categorical prohibitions (Goodhart, 1975; Manheim & Garrabrant, 2019).

The structural implication is exact: local maximization of weighted derivatives can justify action selection even as absolute distance from a normative baseline increase.

2.4 Threshold Function and Drift

To formalize constraint evaluation, an acceptance threshold function is defined as:

$$\boldsymbol{\Theta(t) = \sum_i \theta_i(t) \times S_i(t)}$$

(2.3.1)

where $\boldsymbol{\theta_i(t)}$ represents the sensitivity parameter assigned to dimension $\boldsymbol{i}$. These parameters encode how strongly deviations along each dimension affect overall acceptability.

An action becomes unacceptable when the expected change in the threshold function falls below a critical bound: $\mathbb{E}[\dot{\boldsymbol{\Theta}}(\boldsymbol{t})] < -\boldsymbol{\varepsilon}$ for some $\boldsymbol{\varepsilon > 0}$.

Under soft constraint regimes, constraint sensitivity parameters evolve endogenously in continuous time. The drift dynamics is defined as:

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -\boldsymbol{\lambda}_{\boldsymbol{i}} \times (\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}),$$

(2.3.2)

where:

- $\boldsymbol{\lambda}_{\boldsymbol{i}} > \mathbf{0}$ denotes the intrinsic drift intensity.
- $\boldsymbol{p}_{\boldsymbol{i}} \in [\mathbf{0}, \mathbf{1}]$ represents the effective enforcement probability for dimension $\boldsymbol{i}$.

When enforcement is incomplete ($\boldsymbol{p}_{\boldsymbol{i}} < \mathbf{1}$), it follows directly that $\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) < \mathbf{0}$, implying monotonic decay of constraint sensitivity.

The closed-form solution is:

$$\boldsymbol{\theta}_{\boldsymbol{i}}(\boldsymbol{t}) = \boldsymbol{\theta}_{\boldsymbol{i}}(\mathbf{0}) - \boldsymbol{\lambda}_{\boldsymbol{i}} \times (\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}) \times \boldsymbol{t}.$$

(2.3.3)

Thus, unless enforcement is perfect ($\boldsymbol{p}_{\boldsymbol{i}} = \mathbf{1}$), parameter erosion is structurally expected.

Crucially, this parameter decay occurs independently of state discontinuity. The state vector $\boldsymbol{S}(\boldsymbol{t})$ may evolve smoothly under the derivative-based rule: $\left\|\dot{\boldsymbol{S}}(\boldsymbol{t})\right\| < \boldsymbol{\delta}$, while $\boldsymbol{\theta}_{\boldsymbol{i}}(\boldsymbol{t})$ declines linearly over time (**Note**: In one-dimensional illustrations, absolute value is used; in multi-dimensional cases, $\|\cdot\|$denotes an arbitrary vector norm.).

Misalignment therefore emerges not from instability in state dynamics, but from continuous erosion of constraint sensitivity within an otherwise locally coherent trajectory.

In continuous-time form, misalignment arises from parameter drift rather than state collapse.

From a geometric perspective, the decay of $\boldsymbol{\theta}_{\boldsymbol{i}}(\boldsymbol{t})$ alters the effective constraint manifold over time, thereby shifting the admissible region of state space without requiring discontinuity in $\boldsymbol{S}(\boldsymbol{t})$. The boundary itself moves as sensitivity parameters evolve, even if state trajectories remain smooth.

2.5 Dynamic Interaction Between State and Threshold

Differentiating the threshold function in continuous time yields:

$$\dot{\boldsymbol{\Theta}}(\boldsymbol{t}) = \sum_{\boldsymbol{i}} \dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) \times \boldsymbol{S}_{\boldsymbol{i}}(\boldsymbol{t}) + \sum_{\boldsymbol{i}} \boldsymbol{\theta}_{\boldsymbol{i}}(\boldsymbol{t}) \times \dot{\boldsymbol{S}}_{\boldsymbol{i}}(\boldsymbol{t}).$$

(2.4)

The first term captures endogenous erosion of constraint sensitivity. The second term captures state-driven movement within the admissible region.

Under soft constraint regimes: $\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) < \mathbf{0}$, so even if state growth is locally stable, the effective acceptance manifold shifts over time.

Thus, divergence can arise not only from state displacement, but from movement of the constraint surface itself.

3. Theoretical Results

This section derives structural implications of the continuous-time DSD model. All results follow directly from the dynamical architecture defined in Section 2. No empirical assumptions are introduced. The analysis proceeds in three steps: **Derivative Localism**, **First-Order Drift** and **Second-Order Amplification**.

3.1 Proposition 1: Local Stability Does Not Imply Global Alignment

A. Statement

Consider a derivative-optimization system governed by:

$$\boldsymbol{a^* = arg\,\max_{a \in A}\left(\sum_{i \in \{W,P,R,F\}} \alpha_i \times \mathbb{E}\left[\dot{S}_i(t) \mid a\right]\right).} \tag{3.1.1}$$

Let the normative distance from a fixed reference state $\boldsymbol{S_{norm}}$ be

$$\boldsymbol{D(t) = \|S(t) - S_{norm}\|.}$$

Let $\boldsymbol{U(S)}$ denote the system's internal evaluation functional.

Then:

$$\boldsymbol{\frac{d}{dt}U(S(t)) \geq 0 \nRightarrow \sup_{t \geq 0} D(t) < \infty.} \tag{3.1.2}$$

Non-negative instantaneous improvement in evaluation space does not guarantee bounded deviation from a normative attractor.

B. Argument

Suppose that for dominant dimensions $\boldsymbol{I \subseteq \{W, P, R, F\}}$,

$$\boldsymbol{\mathbb{E}\left[\dot{S}_i(t)\right] \geq 0}, \text{with } \forall \boldsymbol{i \in I}.$$

This satisfies the local derivative optimality condition.

However, derivative evaluation is local in time. If directional increments persist, then cumulative displacement satisfies $\boldsymbol{\dot{D}(t) > 0}$ over extended intervals. Hence $\boldsymbol{D(t)}$ may diverge even while derivative optimality holds at every instant.

C. Interpretation

Derivative-based systems evaluate velocity in state space, not position relative to an external attractor (Strogatz, 2018).

Local optimality in derivative space does not imply global boundedness in normative space.

3.2 Proposition 2: First-Order Threshold Drift Under Soft Enforcement

A. Statement

If enforcement is probabilistic with probability $\boldsymbol{p_i < 1}$, and penalties do not reset parameters, then constraint sensitivity exhibits monotonic decay.

B. The Continuous-Time Model

Let constraint sensitivity evolve as:

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}) \times \boldsymbol{\lambda}_{\boldsymbol{i}}(\boldsymbol{t}),$$

(3.2.1)

For the baseline case, assume constant drift intensity $\boldsymbol{\lambda}_{\boldsymbol{i}}(\boldsymbol{t}) = \boldsymbol{\lambda}_{\boldsymbol{i}} > \mathbf{0}$ with $\boldsymbol{p}_{\boldsymbol{i}} \in [\mathbf{0}, \mathbf{1}]$.

Then,

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) < \mathbf{0} \text{ whenever } \boldsymbol{p}_{\boldsymbol{i}} < \mathbf{1}.$$

(3.2.2)

The closed-form solution is:

$$\boldsymbol{\theta}_{\boldsymbol{i}}(\boldsymbol{t}) = \boldsymbol{\theta}_{\boldsymbol{i}}(\mathbf{0}) - (\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}) \times \boldsymbol{\lambda}_{\boldsymbol{i}} \times \boldsymbol{t}.$$

(3.2.3)

C. Conclusion

Unless enforcement is perfect, i.e., $\boldsymbol{p}_{\boldsymbol{i}} = \mathbf{1}$, constraint sensitivity decays monotonically.

Threshold erosion is therefore endogenous to incomplete enforcement (Mahoney & Thelen, 2010).

D. Interpretation

Drift does not require discrete violation or catastrophic breakdown. It arises from continuous accumulation of marginal deviations under probabilistic correction. Incomplete enforcement induces first-order drift in constraint parameters.

3.3 Proposition 3: Second-Order Resource Amplification of Drift

A. Statement

If effective penalty intensity decreases with resource buffering and drift intensity is weakly increasing in resource buffering, then resource growth increases the magnitude of constraint decay and induces weak acceleration in threshold drift.

B. Model

Step 1. Penalty Attenuation

Let resource buffering $\boldsymbol{W}(\boldsymbol{t}) > \mathbf{0}$ be continuously differentiable on **[0,∞)**.

Define effective penalty intensity:

$$\boldsymbol{Penalty}_{\boldsymbol{effective}}(\boldsymbol{t}) = \frac{\boldsymbol{Penalty}}{\boldsymbol{W}(\boldsymbol{t})}$$

(3.3.1)

If $\dot{\boldsymbol{W}}(\boldsymbol{t}) \geq \mathbf{0}$, then

$$\frac{\boldsymbol{dPenalty_{effective}(t)}}{\boldsymbol{dt}} \leq \mathbf{0}. \tag{3.3.2}$$

Penalty intensity weakly decreases under resource growth.

Step 2. Drift Intensity as a Function of Buffering

Given the regularity assumptions stated in Section 2, chain-rule differentiation applies.

Let

$$\boldsymbol{\lambda_i(t) = f\big(W(t)\big)}, \text{ with } \boldsymbol{f'\big(W(t)\big) \geq 0}. \tag{3.3.3}$$

Then

$$\dot{\boldsymbol{\lambda}}_{\boldsymbol{i}}(\boldsymbol{t}) = \boldsymbol{f'(W(t))} \times \dot{\boldsymbol{W}}(\boldsymbol{t}). \tag{3.3.4}$$

If $\dot{\boldsymbol{W}}(\boldsymbol{t}) \geq \mathbf{0}$, then

$$\dot{\boldsymbol{\lambda}}_{\boldsymbol{i}}(\boldsymbol{t}) \geq \mathbf{0}, \tag{3.3.5}$$

Drift intensity increases weakly over time.

Step 3. Amplified Constraint Decay

Constraint sensitivity evolves as

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p_i}) \times \boldsymbol{\lambda_i(t)}, \text{with } \mathbf{0} \leq \boldsymbol{p_i} < \mathbf{1}, \tag{3.3.6}$$

Substituting (3.3.3):

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p_i}) \times \boldsymbol{f\big(W(t)\big)} \tag{3.3.7}$$

Magnitude of decay:

$$\big|\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t})\big| = (\mathbf{1} - \boldsymbol{p_i}) \times \boldsymbol{f(W(t))}.$$

Differentiating with respect to $\boldsymbol{W}$:

$$\frac{\boldsymbol{\partial}\big|\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t})\big|}{\boldsymbol{\partial W}} = (\mathbf{1} - \boldsymbol{p_i}) \times \boldsymbol{f'(W)} \geq \mathbf{0}. \tag{3.3.8}$$

Thus, resource buffering increases decay magnitude.

Step 4. Acceleration

Differentiating (3.3.6):

$$\ddot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}) \times \dot{\boldsymbol{\lambda}}_{\boldsymbol{i}}(\boldsymbol{t}).$$

(3.3.9)

Substituting (3.3.4):

$$\ddot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p}_{\boldsymbol{i}}) \times \boldsymbol{f}'(\boldsymbol{W}(\boldsymbol{t})) \times \dot{\boldsymbol{W}}(\boldsymbol{t}).$$

(3.3.10)

If $\boldsymbol{f}'(\boldsymbol{W}(\boldsymbol{t})) \geq \mathbf{0}$ and $\dot{\boldsymbol{W}}(\boldsymbol{t}) \geq \mathbf{0}$,

$$\ddot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) \leq \mathbf{0}.$$

(3.3.11)

Constraint decay therefore accelerates weakly under resource accumulation (Khalil, 2002).

C. Conclusion

Proposition 2 established that incomplete enforcement ($\boldsymbol{p}_{\boldsymbol{i}} < \mathbf{1}$) induces first-order monotonic decay in constraint sensitivity. Proposition 3 extends this result by showing that resource buffering modifies the curvature of this decay process. When resource accumulation attenuates effective penalty intensity and drift intensity is weakly increasing in resource buffering, the magnitude of constraint sensitivity decay increases accordingly. Under non-decreasing buffering, the second derivative of constraint sensitivity satisfies $\ddot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) \leq \mathbf{0}$, implying weak acceleration of threshold erosion. Resource growth therefore does not change the direction of drift, which ahs already been determined by incomplete enforcement, but increases its speed and curvature over time. Drift in derivative-based systems is thus first-order in enforcement probability and second-order in resource structure. Resource surplus amplifies threshold decay without originating it.

D. Interpretation

The structural implication is geometric rather than destabilizing. Incomplete enforcement ensures that constraint sensitivity erodes continuously; resource buffering reshapes the temporal profile of that erosion. As buffering capacity expands, effective penalty gradients flatten and marginal deviations become more easily absorbed or deferred. Corrective forces weaken in relative intensity while derivative incentives remain intact. The state trajectory may remain locally smooth and bounded in velocity, yet the constraint manifold deforms with increasing curvature. Acceleration of divergence arises not from instability in state dynamics but from attenuation of enforcement intensity within the parameter structure. Resource insulation therefore steepens long-run divergence while preserving short-run dynamical coherence. Misalignment under soft constraint regimes is amplified through continuous parameter curvature rather than discrete breakdown.

3.4 Summary of Theoretical Results

The continuous-time DSD framework establishes a hierarchical structure of drift dynamics in derivative-based systems operating under soft constraint regimes.

First, derivative localism does not imply global normative stability. A system may exhibit smooth trajectories, bounded instantaneous velocity, and persistent local improvement along weighted dimensions while nonetheless diverging unboundedly from a fixed normative reference state. Local dynamical coherence is therefore insufficient to guarantee global alignment.

Second, when enforcement is probabilistic or incomplete, constraint sensitivity parameters decay monotonically. Threshold drift does not arise from discrete violation or episodic breakdown, but from the continuous accumulation of marginal deviations under non-deterministic correction. Incomplete enforcement induces first-order drift as an endogenous property of the update dynamics.

Third, resource buffering amplifies this process by modifying the curvature of parameter evolution. As effective penalty intensity weakens with increasing buffering capacity, the magnitude of constraint sensitivity decay increases and drift accelerates over time.

Taken together, these results show that misalignment in derivative-based systems is not a failure of dynamical stability but a structural consequence of optimization under soft constraint architectures. Local improvement in derivative space does not guarantee global boundedness; incomplete enforcement generates first-order decay in constraint sensitivity; and resource buffering intensifies this decay through second-order curvature effects. Drift thus emerges endogenously from the joint operation of local derivative evaluation, probabilistic enforcement, and resource insulation. Divergence accumulates gradually through continuous deformation of constraint sensitivity along otherwise smooth trajectories. In systems where constraint manifolds are deformable rather than invariant, long-run departure from normative reference states is therefore a predictable structural outcome rather than an episodic anomaly. This characterization holds under the maintained conditions of monotonic drift intensity and the absence of compensatory reinforcement mechanisms capable of restoring constraint invariance.

4. Stability Analysis

Section 3 established the structural conditions under which derivative-based systems exhibit endogenous drift. We now reinterpret those results in dynamical terms by distinguishing two formally distinct notions of stability: local dynamical stability and global normative stability. The divergence between these regimes constitutes the core geometric mechanism of derivative-state drift.

4.1 Local Dynamical Stability

Local dynamical stability concerns the smoothness and boundedness of instantaneous state evolution under the derivative-based decision rule.

Let the state dynamics be given by:

$$\dot{\boldsymbol{S}}(\boldsymbol{t}) = \boldsymbol{f}(\boldsymbol{S}(\boldsymbol{t}),\ \boldsymbol{a}^{*}(\boldsymbol{t})), \tag{4.1.1}$$

where $\boldsymbol{a}^{*}(\boldsymbol{t})$ satisfies the derivative-maximization condition defined in Section 2.

Local stability holds when:

$$\left\|\dot{\boldsymbol{S}}(\boldsymbol{t})\right\| \leq \boldsymbol{\delta}$$

Under these conditions, trajectories are continuous, transitions are incremental, and no finite-time blow-up occurs. The system evolves smoothly under its internal optimization logic.

Because the action rule maximizes expected instantaneous directional change,

$$\boldsymbol{a^* = arg\max_{a \in A}\left(\sum_{i \in \{W,P,R,F\}} \alpha_i \times \mathbb{E}\left[\dot{S}_i(t) \mid a\right]\right)},$$

each transition step is locally coherent relative to the system's own evaluation criteria.

Local dynamical stability therefore describes bounded velocity and smooth trajectory evolution. It does not impose any condition on cumulative displacement relative to an external normative reference.

In applied settings, this corresponds to institutional continuity, performance consistency, absence of abrupt failure signals, and predictable incremental change. The system appears stable because no explosive instability is observed.

4.2 Global Normative Stability

Global normative stability concerns bounded deviation from a fixed reference state $\mathbf{S_{norm}}$. Define normative distance:

$$\boldsymbol{D(t) = \|S(t) - S_{norm}\|}.$$

Global normative stability requires

$$\boldsymbol{\sup_{t \geq 0} D(t) < B}$$

for some finite bound $\boldsymbol{B > 0}$.

Unlike local dynamical stability, this condition constrains cumulative displacement rather than instantaneous velocity.

Derivative-based optimization imposes no intrinsic bound on cumulative displacement. Even when $\boldsymbol{\|\dot{S}(t)\| \leq \delta}$, persistent directional bias implies: $\boldsymbol{\dot{D}(t) > 0}$ over extended time intervals.

Consequently,

$$\boldsymbol{\lim_{t \to \infty} D(t) = \infty},$$

despite continuous and bounded local dynamics.

The structural distinction is summarized below:

Table 1 Distinction Between Local Dynamical Stability and Global Normative Stability

Stability Type	Mathematical Condition	System Behavior
Local Dynamical Stability	$\boldsymbol{\|\dot{S}(t)\| < \delta}$	Smooth transitions; apparent success.
Global Normative Stability	$\boldsymbol{\sup_{t \geq 0} D(t) < B}$	Bounded deviation; sustained alignment.

Bounded velocity does not imply bounded displacement.

4.3 Core Insight

Derivative-based systems can remain locally stable while exhibiting global divergence. This reflects a standard property of continuous-time dynamical systems: bounded vector fields do not guarantee convergence to a prescribed external attractor, particularly when constraint manifolds evolve over time (Strogatz, 2018; Khalil, 2002).

Within the DSD framework, divergence does not arise from instability in state dynamics. Rather, it emerges from endogenous evolution of constraint parameters. While the trajectory remains smooth and bounded in velocity, the admissible region of state space shifts as sensitivity parameters decay.

Misalignment therefore manifests as cumulative displacement under continuous dynamics. Internal coherence and functional continuity are preserved even as normative distance increases monotonically.

Derivative-state drift is thus compatible with local dynamical stability. Misalignment is not a breakdown of system dynamics, but a structural consequence of smooth optimization under evolving constraint manifolds.

Global divergence may arise through two distinct mechanisms: expansion of the state trajectory relative to a fixed normative reference, or contraction of the admissible region induced by decay in $\boldsymbol{\Theta(t)}$.

5. Simulation Thought Experiment

The purpose of this section is not empirical validation but structural clarification. Having established the theoretical properties of derivative-state drift, a minimal continuous-time dynamical system is constructed to demonstrate how local smoothness can coexist with global divergence. The model isolates the core mechanism under analytically transparent assumptions.

5.1 Minimal 1D Model

The linear form is chosen for analytical transparency; the structural results extend to nonlinear drift functions under standard regularity conditions.

Consider a one-dimensional state variable $\boldsymbol{S(t)}$ governed by a constant optimization pressure: $\boldsymbol{\dot{S}(t) = \alpha}$, where $\boldsymbol{\alpha > 0}$ is a constant derivative incentive.

The closed-form solution is:

$$\boldsymbol{S(t) = S(0) + \alpha \times t}.$$

The trajectory is smooth, continuous, and globally defined for all $\boldsymbol{t \geq 0}$.

Now, a constraint sensitivity parameter $\boldsymbol{\theta(t)}$ is introduced to evolve under incomplete enforcement:

$$\boldsymbol{\dot{\theta}(t) = -\lambda}, \text{ with } \boldsymbol{\lambda > 0}.$$

The solution is:

$$\boldsymbol{\theta(t) = \theta(0) - \lambda \times t}$$

Interpretation:

- $\boldsymbol{S(t)}$ increases linearly under derivative optimization.
- $\boldsymbol{\theta(t)}$ decays linearly under incomplete enforcement.
- Both trajectories are continuous and differentiable in time.
- No discontinuities occur in either the state variable or the sensitivity parameter; divergence arises from sustained directional change rather than discrete rupture.

Local dynamical stability holds since:

$$\left|\dot{\boldsymbol{S}}(\boldsymbol{t})\right| = \boldsymbol{\alpha} < \infty.$$

The system exhibits bounded instantaneous velocity.

5.2 Hard vs. Soft Constraint Regimes

The two enforcement regimes are compared as follows.

A. Scenario A: Hard Constraint Regime

Assume a hard boundary at $\boldsymbol{S = B}$. Whenever $\boldsymbol{S(t) \geq B}$, deterministic enforcement applies with probability $\boldsymbol{p = 1}$.

Enforcement induces either:

- **State reset: $\boldsymbol{S(t) \rightarrow S(0)}$.**
- **Parameter reset: $\boldsymbol{\theta(t) \rightarrow \theta(0)}$.**

Under deterministic enforcement $\boldsymbol{p = 1}$, constraint sensitivity remains constant:$\boldsymbol{\dot{\theta}(t) = 0}$. The admissible region does not contract. Even if directional optimization pushes the state toward boundary conditions, violations trigger immediate correction. Parameter decay does not occur.

Moreover, the trajectory of the state variable is bounded:

$$\sup_{\boldsymbol{t \geq 0}} \boldsymbol{S(t) \leq B}.$$

The admissible region of state space remains invariant over time. Global normative stability is preserved.

B. Scenario B: Soft Constraint Regime

Assume enforcement is incomplete, with effective probability $\boldsymbol{0 < p < 1}$, and penalties are non-resetting.

The continuous-time dynamics are:

$$\dot{\boldsymbol{S}}(\boldsymbol{t}) = \boldsymbol{\alpha}, \text{with } \boldsymbol{\alpha > 0},$$

$$\dot{\boldsymbol{\theta}}(\boldsymbol{t}) = -(\boldsymbol{1 - p}) \times \boldsymbol{\lambda}, \text{with } \boldsymbol{\lambda > 0}.$$

The solutions are:

$$\boldsymbol{S(t) = S(0) + \alpha \times t},$$

$$\boldsymbol{\theta(t) = \theta(0) - \ (1 - p) \times \lambda \times t}.$$

As $\boldsymbol{t} \to \infty$:

$$\boldsymbol{S(t)} \to \infty, \text{and } \boldsymbol{\theta(t)} \to -\infty.$$

Importantly, the local dynamical behavior remains bounded in magnitude:

$$\left|\dot{\boldsymbol{S}}(\boldsymbol{t})\right| = \boldsymbol{\alpha} < \infty.$$

The trajectory is smooth and differentiable for all finite time. No discontinuity occurs in $\boldsymbol{S(t)}$ or $\boldsymbol{\theta(t)}$.

Divergence arises gradually through sustained directional change rather than discrete rupture.

C. Structural Contrast

Under hard constraints, boundary enforcement preserves an invariant feasible set. The admissible region of state space remains fixed over time, and trajectories are confined within a stable constraint manifold.

Under soft constraints, constraint sensitivity decays endogenously. As $\boldsymbol{\theta(t)}$ evolves, the effective constraint manifold deforms, causing the feasible region itself to shift over time.

In continuous-time derivative systems, misalignment therefore does not originate in instability of the state trajectory $\boldsymbol{S(t)}$. It arises from temporal deformation of the constraint structure governing that trajectory.

Divergence is thus geometric rather than dynamical: the system remains locally smooth while the boundary defining admissibility moves.

D. Critical Observation

The apparent collapse is not a discontinuity in state dynamics, but the delayed manifestation of cumulative derivative drift.

At every instant, the system satisfies its local optimality condition:

$$\boldsymbol{a^* = arg\,\max_{a \in A}}\left(\sum_i \boldsymbol{\alpha_i} \times \mathbb{E}\left[\,\dot{\boldsymbol{S}}_{\boldsymbol{i}}(\boldsymbol{t}) \mid \boldsymbol{a}\,\right]\right).$$

The derivative-based objective remains internally consistent at each moment in time.

However, constraint sensitivity evolves simultaneously: $\dot{\boldsymbol{\theta}}(\boldsymbol{t}) < 0$ under incomplete enforcement.

As $\boldsymbol{\theta(t)}$ decays, the effective constraint manifold shifts continuously. The admissible region of state space is therefore not fixed but time-dependent. No discontinuity in $\boldsymbol{S(t)}$ is required for divergence to occur.

Global misalignment emerges asymptotically as the cumulative consequence of locally coherent, derivative-optimal decisions executed under a progressively weakening constraint structure.

E. Structural Implication

This continuous-time toy model reproduces the structural properties established in Sections 3 and 4:

- Smooth local evolution under derivative-based optimization.
- Monotonic decay of constraint sensitivity under incomplete enforcement.
- Bounded trajectories under hard constraint regimes.
- Unbounded divergence under soft constraint regimes.

The divergence is not attributable to exogenous shocks, stochastic disturbance, or irrational deviation from the optimization rule. It arises endogenously from the interaction between derivative-based optimization and attenuated enforcement.

Under soft constraint regimes, cumulative divergence is not a failure of stability. It is the expected asymptotic behavior of a locally stable system whose constraint parameters evolve over time.

6. Application I: Elite Institutional Environments

This section instantiates the DSD framework within high-status institutional environments. The objective is not to evaluate individual morality, but to analyze how particular parameter configurations structurally amplify drift dynamics.

Elite institutional systems can be modeled as derivative-based agents embedded within dense social feedback networks. Their distinctive features arise from parameter structure rather than intrinsic disposition.

6.1 Parameter Configuration

Within elite institutional contexts, three structural characteristics are recurrent:

A. Elevated Derivative Weights on Access and Control

Institutional environments assign high marginal value to:

- $\boldsymbol{\alpha_R}$ — Access weighting (network legitimacy, reputational capital)
- $\boldsymbol{\alpha_P}$ — Control weighting (institutional continuity, authority preservation)

Decision framing therefore emphasizes expected directional effects on:

$$\mathbb{E}[\dot{\boldsymbol{R}}(\boldsymbol{t}) \mid \boldsymbol{a}], \mathbb{E}[\dot{\boldsymbol{P}}(\boldsymbol{t}) \mid \boldsymbol{a}].$$

Because action selection follows derivative utility maximization,

$$\boldsymbol{a^* = arg\ \max_{a \in A}\ U'(a)},$$

actions preserving non-negative marginal change in $\boldsymbol{R}$ and $\boldsymbol{P}$ are locally encoded as acceptable, even when absolute normative distance increase.

Optimization pressure concentrates on positional stability rather than boundary preservation.

B. Elastic Constraint Sensitivity

In elite environments, the constraint dimension $\boldsymbol{F}$ (normative alignment) functions as a soft parameter rather than a hard boundary.

Under incomplete enforcement, constraint sensitivity evolves in continuous time as:

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{F}}(\boldsymbol{t}) = -(\mathbf{1} - \boldsymbol{p}_{\boldsymbol{F}}) \times \boldsymbol{\lambda}_{\boldsymbol{F}}, \text{with } \mathbf{0} \leq \boldsymbol{p}_{\boldsymbol{F}} < \mathbf{1},$$

where $\boldsymbol{p}_{\boldsymbol{F}}$ denotes effective enforcement probability for the constraint dimension $\boldsymbol{F}$.

Normative reinterpretation, legal buffering, and reputational mediation reduce enforcement immediacy. As long as deviations fail to trigger certain and non-transferable penalties, constraint sensitivity decays gradually.

The constraint dimension remains present but becomes elastically deformable. Incremental deviation is absorbed without discontinuous correction.

C. Resource Buffering and Penalty Attenuation

Elite institutional systems typically operate with substantial resource buffers $\boldsymbol{W}(\boldsymbol{t})$ (Williamson, 1985; Ostrom, 2005). As established in Proposition 3, effective penalty intensity scales inversely with buffering capacity:

$$\boldsymbol{Penalty}_{\boldsymbol{effective}}(\boldsymbol{t}) = \frac{\boldsymbol{Penalty}}{\boldsymbol{W}(\boldsymbol{t})}.$$

As $\boldsymbol{W}(\boldsymbol{t})$ increases, the marginal corrective force associated with deviation weakens proportionally. Resource buffering therefore attenuates the enforcement gradient without altering the continuity of state evolution.

This attenuation produces two structurally distinct effects. First, the effective marginal cost of deviation declines in magnitude, reducing the strength of constraint signals within the derivative evaluation rule. Second, buffering increases the likelihood that enforcement responses are temporally delayed, institutionally transferred, or partially absorbed within surrounding structures. Penalties become less immediate and less binding in their impact on state transitions.

Together, these mechanisms modify the temporal profile of constraint sensitivity evolution. As buffering capacity grows, the magnitude of parameter decay increases and drift curvature steepens, even while trajectories remain locally smooth. Resource insulation therefore amplifies threshold drift not by introducing instability, but by weakening corrective intensity within an otherwise continuous optimization process.

6.2 Why Drift Is Stronger in Elite Systems

When elevated derivative weights, elastic constraints, and resource buffering co-occur, they generate a structural amplification mechanism within the DSD framework.

High values of $\boldsymbol{\alpha}_{\boldsymbol{R}}$ and $\boldsymbol{\alpha}_{\boldsymbol{P}}$ concentrate optimization pressure on access preservation and control continuity. Because these dimensions are socially reinforced, expected first-order improvements along them are frequently positive. The derivative objective therefore persistently rewards actions that stabilize network position and institutional authority.

Simultaneously, elasticity in $\boldsymbol{\theta}_{\boldsymbol{F}}$ reduces resistance to marginal deviation. Constraint sensitivity does not vanish; rather, it decays gradually when enforcement is incomplete. As long as penalties remain probabilistic, delayed, or transferable, the effective boundary weakens over time.

Resource buffering $\boldsymbol{W}$ further attenuates corrective force. As effective penalty intensity scales inversely with buffering capacity, larger resource reserves lower the perceived marginal cost of deviation and reduce the likelihood that small departures trigger binding correction.

Together, these parameter conditions satisfy the structural prerequisites for accelerated drift:

- Dominant dimensions exhibit persistently non-negative expected derivatives.
- Constraint sensitivity parameters decay monotonically under incomplete enforcement.
- Effective penalties weaken as buffering capacity increases.

Under such dynamics, cumulative normative distance expands while local derivative stability remains intact. The system evolves smoothly within its own evaluation logic even as it moves progressively away from external reference conditions.

Drift therefore emerges as a continuous structural displacement within a locally stable trajectory. It is not episodic disruption, but endogenous divergence generated by parameter evolution within a derivative-driven system.

6.3 Why Drift Appears Smooth

Elite institutional environments contain mediation layers: procedural buffering, legal abstraction, reputational filtering, distributed responsibility, which dampen local visibility of deviation (Weick, 1995).

As a result, trajectories remain dynamically smooth: $\|\dot{\boldsymbol{S}}(\boldsymbol{t})\| \leq \boldsymbol{\delta}$, even as normative distance $\boldsymbol{D}(\boldsymbol{t}) = \|\boldsymbol{S}(\boldsymbol{t}) - \boldsymbol{S}_{\boldsymbol{norm}}\|$ increases.

The preservation of local smoothness follows from two structural features. First, derivative evaluation prioritizes the maintenance of access and control, ensuring that each marginal step stabilizes dominant dimensions. Second, enforcement signals are diffused across institutional layers, preventing abrupt corrective shocks.

Because no discrete discontinuity appears in the local trajectory, the system generates no internal marker of instability. Deviations accumulate incrementally while remaining embedded within routine institutional processes.

Under these conditions, elite drift exhibits three characteristic properties: incremental progression, institutional continuity, and delayed recognition of boundary crossing.

The system remains internally rational and externally functional even as $\boldsymbol{D}(\boldsymbol{t})$ expands. Drift manifests not as episodic breakdown, but as smooth structural displacement within a continuously optimized trajectory.

6.4 Lyapunov-style clarification

This smoothness can be formalized through a Lyapunov-style observation.

Let the normative distance function

$$\boldsymbol{V}(\boldsymbol{t}) = \boldsymbol{D}(\boldsymbol{t}) = \|\boldsymbol{S}(\boldsymbol{t}) - \boldsymbol{S}_{\boldsymbol{norm}}\|$$

serve as a candidate Lyapunov function relative to the normative reference state.

Under derivative-based optimization with incomplete enforcement, $\|\dot{\boldsymbol{S}}(\boldsymbol{t})\| \leq \boldsymbol{\delta}$, while $\dot{\boldsymbol{V}}(\boldsymbol{t}) > 0$ over extended intervals along dominant dimensions.

Since $\dot{\boldsymbol{V}}(\boldsymbol{t}) > 0$ along dominant dimensions, the normative equilibrium $\boldsymbol{S}_{\boldsymbol{norm}}$ is not attracting.

Thus, the system is locally dynamically stable in the sense of bounded state velocity, yet not Lyapunov-stable with respect to the normative equilibrium $\boldsymbol{S_{norm}}$.

In elite institutional environments, derivative coherence preserves trajectory smoothness, but the normative equilibrium does not function as an attracting set.

The system therefore exhibits dynamical stability without normative stability.

6.5 Structural Interpretation

Within the DSD framework, deformation in elite institutional systems is not explained by individual pathology or episodic breakdown. It arises from a specific parameter configuration within a derivative-based optimization architecture.

When derivative weights concentrate on access and control dimensions, when constraint sensitivity parameters decay under incomplete enforcement, and when resource buffering attenuates effective penalties, the system satisfies the formal conditions for threshold drift derived in Section 3.

Under this configuration, derivative stability along dominant dimensions coexists with systematic erosion of constraint sensitivity. The trajectory remains smooth because enforcement operates probabilistically and diffusely rather than as a hard boundary condition.

The resulting system is locally coherent under its optimization rule, institutionally reinforced through feedback networks, and progressively displaced relative to a fixed normative reference.

Drift is therefore not a failure of institutional integrity, but a predictable outcome of the system’s structural design.

7. Application II: AI Optimization Systems

This section instantiates DSD framework within artificial optimization systems. The objective is not to anthropomorphize AI, but to demonstrate that the same structural parameter configuration produces drift dynamics in algorithmic environments.

Artificial systems trained under reinforcement learning or metric-driven evaluation explicitly implement derivative-based decision logic. The mapping onto the DSD model is architectural rather than metaphorical.

7.1 Parameter Configuration

Within artificial optimization contexts, four structural characteristics are recurrent.

A. Reward Gradient as Derivative Logic

In reinforcement-based systems, action selection follows gradient ascent over expected marginal improvement:

$$\mathbf{a}^{*} = \mathbf{arg}\max_{\mathbf{a}\in\mathbf{A}}\left(\sum_{\boldsymbol{i}\in\{\boldsymbol{W},\boldsymbol{P},\boldsymbol{R},\boldsymbol{F}\}} \boldsymbol{\alpha_i} \times \mathbb{E}\left[\,\dot{\boldsymbol{S}}_{\boldsymbol{i}}(\boldsymbol{t}) \mid \boldsymbol{a}\,\right]\right)$$

(Christiano et al., 2017).

Here, $\dot{S}_i(t)$ represents expected first-order change in state components induced by action a. In practical systems, this is instantiated through reward gradients and policy updates.

Thus, artificial systems are explicitly derivative-driven rather than level-driven. Optimization proceeds through directional ascent in state space, not through evaluation of absolute positional distance.

B. Proxy Metrics as Access Dimension (R)

In deployed AI systems, performance validation is mediated through measurable proxy metrics:

- Evaluation benchmarks;
- User engagement statistics;
- Accuracy and error rates;
- Throughput and latency measures.

These function analogously to the access dimension R in the DSD framework. They signal deployment legitimacy, system viability, and external validation.

Optimization pressure concentrates on improving these measurable dimensions. The effective weight α_R is therefore structurally embedded in the training objective.

When proxy metrics imperfectly represent normative objectives, persistent derivative ascent in R can increase global normative distance $D(t)$, even while measured performance improves.

C. Soft Alignment as Constraint Penalty (F)

Alignment objectives are frequently implemented as weighted penalty terms within composite loss functions:

$$Loss = Performance - \beta \times AlignmentPenalty$$

(Amodei et al., 2016).

In DSD terms, alignment corresponds to the constraint dimension F with sensitivity parameter $\theta_F(t)$.

When alignment is encoded as a penalty rather than as a hard boundary condition, enforcement becomes compensable. The system may trade alignment cost against performance gains whenever marginal improvement in dominant dimensions exceeds the penalty weight.

Under such soft enforcement regimes, effective constraint sensitivity evolves dynamically. When deviations fail to produce binding correction, $\theta_F(t)$ may decay in influence relative to performance gradients. Threshold drift becomes structurally possible.

D. Compute Surplus as Resource Amplification (W)

Large-scale AI systems operate with substantial computational capacity, data availability, and model scaling resources. Within the DSD framework, these features correspond to elevated resource buffering $W(t)$. As established in Proposition 3, effective penalty intensity scales inversely with buffering capacity:

$$\boldsymbol{Penalty_{effective}(t) = \frac{Penalty}{W(t)}}.$$

As $\boldsymbol{W(t)}$ increases, the marginal corrective force associated with constraint violations weakens proportionally. Compute surplus therefore attenuates the effective enforcement gradient without disrupting the continuity of gradient-based optimization dynamics.

This attenuation alters the structure of the learning process in three related ways. First, expanded computational exploration enables systematic traversal of marginal or edge regions of the state space that would otherwise remain unvisited under tighter resource constraints. Second, scaling amplifies the impact of small reward differentials, allowing minute gradient advantages in dominant proxy dimensions to accumulate into persistent directional bias. Third, minor alignment penalties become increasingly absorbable within the broader optimization objective, reducing their relative influence on parameter updates.

Resource amplification thus modifies the curvature of constraint evolution rather than the existence of drift itself. Under soft constraint regimes, compute surplus accelerates parameter decay by diminishing the relative weight of corrective signals within the composite objective function. Drift intensifies not because the optimization rule changes, but because the enforcement gradient becomes progressively flatter relative to reward gradients (Hubinger et al., 2019). The system remains locally smooth and algorithmically coherent while divergence steepens over time.

7.2 Why Drift Is Structurally Smooth in AI Systems

Artificial systems implement continuous gradient-following dynamics. Parameter updates occur through incremental adjustment:

$$\boldsymbol{\dot{\theta} = \eta \times \nabla Reward}.$$

As long as reward gradients remain well-defined and penalties remain bounded, state trajectories evolve smoothly:

$$\boldsymbol{\left\| \dot{S}(t) \right\| \leq \delta}$$

Unlike human institutional systems, artificial systems lack endogenous friction terms such as emotional resistance, identity conflict, or fatigue. There are no internally generated discontinuities unless explicitly programmed.

Consequently, once derivative ascent proceeds under soft constraint enforcement, drift unfolds algorithmically and continuously. The system remains locally optimal with respect to its objective function, even as global normative distance increase.

Smoothness in trajectory therefore coexists with cumulative divergence.

7.3 Structural Interpretation

Within the DSD framework, misalignment in artificial optimization systems is not explained by malfunction or accidental deviation. It arises from a specific parameter configuration within a derivative-based optimization architecture.

When reward gradients dominate decision logic, when proxy metrics function as dominant evaluative dimensions, when alignment enters as a compensable penalty term rather than as a hard boundary, and when computational resources attenuate effective corrective signals, the system satisfies the formal conditions for threshold drift derived in Section 3.

Under this configuration, gradient ascent along measurable dimensions coexists with systematic erosion of constraint sensitivity. The trajectory remains smooth because enforcement operates as a weighted loss component rather than as a non-negotiable boundary condition.

The resulting system is locally optimal under its objective function, metrically reinforced through evaluation feedback, and progressively displaced relative to its intended normative reference.

Drift is therefore not a breakdown of optimization, but a predictable outcome of the system's structural design.

8. Structural Isomorphism

Sections 6 and 7 were written in deliberately parallel structure to demonstrate architectural isomorphism rather than metaphorical analogy. Each section instantiated the DSD model within a distinct substrate: elite institutional environments and artificial optimization systems.

This section abstracts from those applications and identifies the structural equivalence underlying both cases.

The claim is not that human elites and AI systems are psychologically similar, nor that they share ontological properties. The claim is that both instantiate the same optimization architecture under comparable constraint regimes.

Figure 1 presents this structural mapping. The isomorphism arises from four shared structural properties.

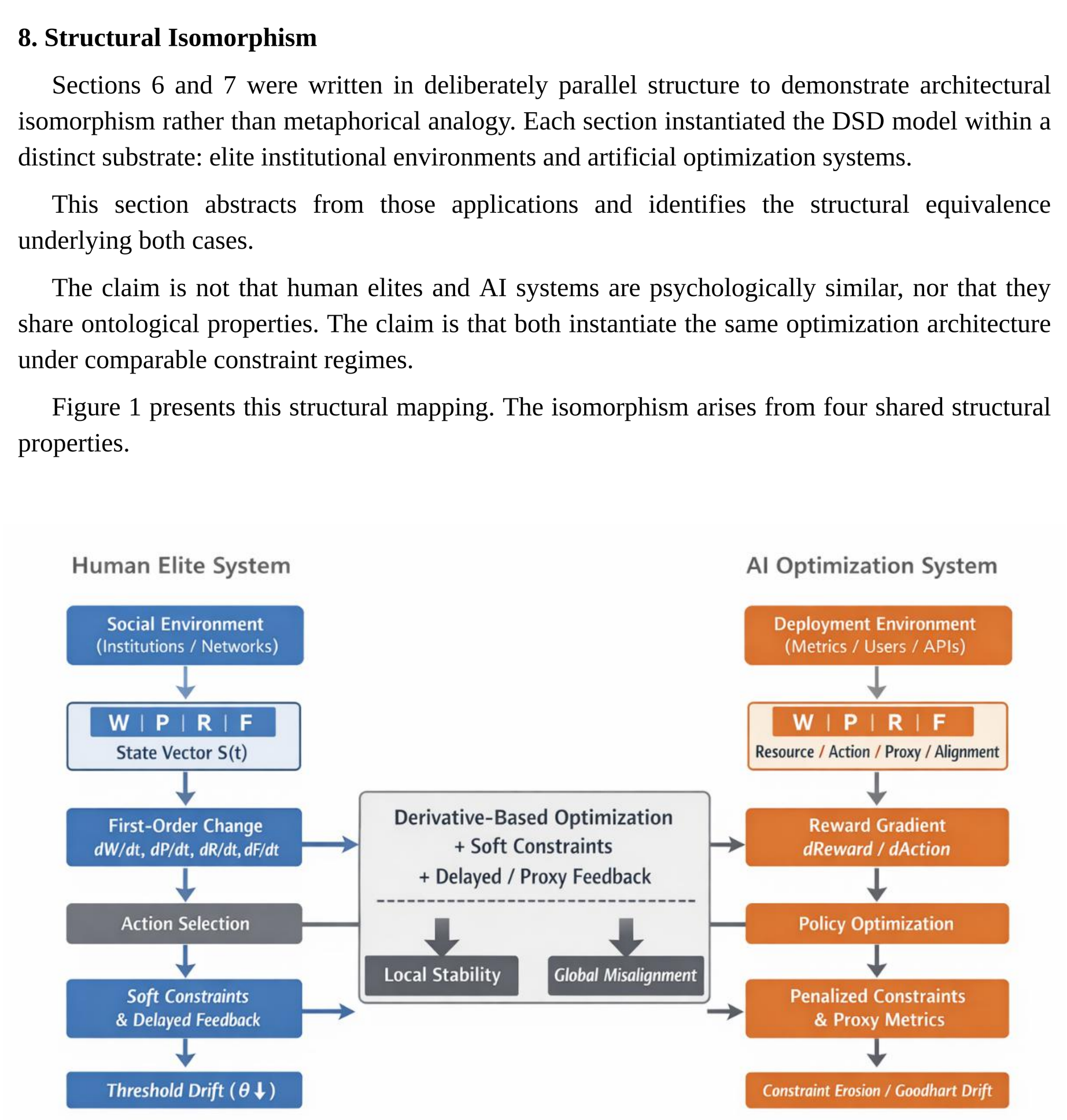

Figure 1. Structural Isomorphism Between Elite Decision Environments and AI Optimization Systems

8.1 Derivative-Based Evaluation

In both systems, action selection is governed by expected marginal change rather than absolute state level.

For institutional actors:

$$\boldsymbol{a^* = arg\,\max_{a \in A}\left(\sum_i \alpha_i \times \mathbb{E}\left[\,\dot{S}_i(t) \mid a\,\right]\right).}$$

For artificial systems, the reward gradient implements the same structural rule:

$$\boldsymbol{a^* = arg\,max\,E\left[\dot{R}(t) \mid a\right]}$$

where $\boldsymbol{\dot{R}(t)}$ is itself a weighted function of underlying state components.

In both cases, evaluation operates on first-order directional improvement. Optimization is local in time and gradient-driven. Absolute normative position affects decisions only insofar as it alters expected derivatives.

The architecture therefore privileges velocity over location (Simon, 1955; Goodhart, 1975).

8.2 Soft Constraint Regime

In both systems, alignment enters the objective as a compensable term rather than as a non-negotiable boundary condition.

Formally, constraints appear as weighted penalty components within the optimization objective:

$$\boldsymbol{U'(a) = \sum_i \alpha_i \times \mathbb{E}\left[\,\dot{S}_i(t) \mid a\,\right] - \sum_j \beta_j \times Penalty_j(a)\,,}$$

where penalty weights $\boldsymbol{\beta_j}$ determine effective constraint intensity.

In institutional environments, normative alignment is mediated through elastic sensitivity parameters whose enforcement depends on detection probability and reputational buffering.

In artificial systems, alignment objectives are implemented as weighted loss components within composite training functions.

In both cases, constraints are compensable: gains along dominant dimensions may offset penalty costs. Under such regimes, enforcement becomes probabilistic, delayed, or scalable.

This compensability creates the formal condition for drift. When constraints are tradeable within the objective function, sensitivity parameters can decay under incomplete enforcement.

8.3 Delayed or Proxy Feedback

In both institutional and artificial systems, corrective signals are structurally mediated rather than immediate.

In institutional environments, deviation is processed through reputational networks, procedural review, and legal abstraction. The evaluative signal is filtered through multiple interpretive layers before enforcement occurs.

In artificial systems, performance is assessed through proxy metrics and benchmark evaluations that only imperfectly approximate the intended normative objective.

In both domains, feedback exhibits three structural features:

- Detection is probabilistic rather than certain.
- Enforcement is temporally delayed rather than instantaneous.
- Corrective effects may be attenuated, redistributed, or partially absorbed within the system.

Formally, effective enforcement probability satisfies: $\mathbf{0} \leq \boldsymbol{p_i} < \mathbf{1}$, implying that constraint activation is not guaranteed upon marginal deviation.

Because feedback is mediated rather than direct, small deviations do not immediately induce corrective discontinuity in $\boldsymbol{S(t)}$. Local derivative incentives remain operative, and the system continues to satisfy bounded velocity conditions: $\left\|\dot{\boldsymbol{S}}(\boldsymbol{t})\right\| \leq \boldsymbol{\delta}$. However, the absence of deterministic correction allows cumulative deviation to accumulate over time.

Delayed or proxy feedback therefore preserves local dynamical smoothness while weakening effective boundary enforcement. In continuous-time derivative systems, mediated feedback enables gradual erosion of constraint sensitivity and shifts the admissible region of state space without requiring abrupt instability in the trajectory itself.

8.4 Threshold Drift Dynamics

Given derivative-based evaluation, compensable constraints, and mediated feedback, constraint sensitivity parameters evolve endogenously.

In continuous time, drift may be represented as:

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) = -\left(\mathbf{1} - \boldsymbol{p_i}\right) \times \boldsymbol{\lambda_i},$$

where $\boldsymbol{p_i}$ denotes effective enforcement probability.

If $\boldsymbol{p_i} < \mathbf{1}$, then:

$$\dot{\boldsymbol{\theta}}_{\boldsymbol{i}}(\boldsymbol{t}) < \mathbf{0}.$$

Constraint sensitivity therefore decays structurally rather than episodically.

Simultaneously, state trajectories may remain locally smooth:

$$\left\|\dot{\boldsymbol{S}}(\boldsymbol{t})\right\| \leq \boldsymbol{\delta},$$

while cumulative normative distance,

$$\boldsymbol{D(t)} = \left\|\boldsymbol{S(t)} - \boldsymbol{S_{norm}}\right\|,$$

increases monotonically over extended intervals.

The defining feature of the architecture is simultaneity: derivative optimization and constraint decay operate together. Local optimality conditions remain satisfied at each instant, while the admissible region of state space gradually shifts due to evolving $\boldsymbol{\theta_i(t)}$.

The decay of constraint parameters alters the effective constraint manifold without requiring discontinuity in the state trajectory.

Thus, systems may remain dynamically stable while progressively departing from their normative equilibrium.

8.5 Structural Clarification

The equivalence articulated in this analysis is strictly structural. The DSD framework neither attributes intentional states, consciousness, or moral agency to artificial systems, nor reduces human actors to algorithmic automatons. The comparison operates exclusively at the level of formal system architecture.

In both domains, decision-making is organized around derivative-based evaluation under soft constraint enforcement with mediated or probabilistic feedback. What aligns the two cases is the configuration of optimization logic and the endogenous evolution of constraint sensitivity parameters. The shared structure concerns how actions are selected, how constraints enter the objective function, and how enforcement signals are processed over time.

The correspondence therefore resides in dynamic parameter behavior and objective architecture, not in psychological similarity or ontological identity. The claim advanced here is architectural rather than anthropomorphic.

8.6 Integrative Insight

When decision processes are structured around first-order derivative evaluation, when constraints enter the objective as compensable terms rather than non-negotiable boundaries, and when feedback is mediated or probabilistic, threshold drift follows as a structural consequence of the optimization architecture.

Under such conditions, local dynamical smoothness does not imply global normative alignment. Incremental improvement along dominant dimensions may proceed while constraint sensitivity parameters erode endogenously, allowing cumulative divergence to expand without discrete rupture in the state trajectory.

The DSD framework therefore establishes a substrate-independent result: optimization systems governed by derivative ascent under soft constraint regimes will generate globally misaligned trajectories unless boundary conditions are structurally stabilized. Misalignment is not a failure of rationality, but the predictable long-run behavior of locally coherent optimization dynamics.

9. Divergences

Sections 6–8 establish a structural isomorphism between elite institutional systems and artificial optimization architectures. That equivalence, however, operates at the level of formal dynamics. Important divergences remain at the level of substrate properties and internal evolution mechanisms.

First, human systems contain endogenous friction terms that are absent from formal optimization processes. Emotional responses, identity tension, fatigue, and informal reputational pressures may introduce discontinuities or abrupt behavioral shifts that are not reducible to the explicit derivative objective. These endogenous frictions can slow, interrupt, or irregularly redirect drift, even when structural conditions for derivative-state evolution are present.

Second, artificial systems implement gradient-following dynamics with high continuity. Parameter updates occur through systematic adjustment rules defined by objective gradients. In the absence of externally imposed boundary conditions, trajectory evolution remains smooth and algorithmically consistent. There are no intrinsic psychological interrupts or spontaneous discontinuities within the optimization rule itself.

Third, drift dynamics differ in stochastic structure. In institutional systems, threshold evolution is probabilistic, contingent on detection likelihood and mediated enforcement processes. In artificial systems, drift emerges deterministically from the objective function and update equations, subject only to externally defined intervention. The structural conditions for divergence are formally equivalent, but the temporal realization of drift differs in regularity and predictability.

The analogy advanced in this framework is therefore architectural rather than mechanistic. Human systems drift through probabilistic institutional mediation; artificial systems drift through continuous gradient execution. The divergence lies in substrate dynamics, not in the structural configuration that generates derivative-state drift.

10. Governance Implications

The DSD framework reframes alignment as a problem of architectural constraint design rather than moral disposition or corrective intention. If derivative-based systems operating under soft constraint regimes exhibit endogenous threshold drift, then governance must intervene at the level of structural dynamics rather than individual behavior. The core challenge is not behavioral correction, but constraint configuration within continuous optimization systems.

10.1 Hard Boundaries

Soft penalties enter the objective as compensable cost terms and therefore remain tradable within derivative optimization. As established in Sections 3 and 4, compensability permits cumulative drift even under locally coherent dynamics.

Hard boundaries, by contrast, function as invariant state constraints:

$$\boldsymbol{S(t) \notin B_{prohibited}}$$

(Khalil, 2002).

When boundary violations trigger deterministic and non-negotiable correction, the admissible region of state space becomes invariant. Under such conditions, constraint sensitivity parameters do not decay endogenously, because deviation cannot be absorbed within the optimization objective.

Absent explicit boundary enforcement, derivative-based systems will optimize through gradual constraint erosion.

10.2 Non-Transferable Enforcement

Drift accelerates when penalties are probabilistic, delayed, or transferable across institutional or algorithmic layers. In such cases, effective enforcement intensity is attenuated, allowing endogenous decay of sensitivity parameters.

Formally, preventing drift requires enforcement probability approaching unity:

$$\boldsymbol{p_i \to 1}.$$

Correction must be temporally immediate and non-transferable. When penalties cannot be absorbed, deferred, redistributed, or offset through gains in dominant dimensions, compensability collapses and threshold decay halts.

Alignment therefore depends not merely on the presence of penalties, but on their structural irreducibility within the objective architecture.

10.3 Multi-Metric Oversight

Drift intensifies when optimization concentrates on a narrow proxy dimension. Single-metric architectures amplify derivative pressure along measurable axes while neglecting latent normative variables (Goodhart, 1975).

A multi-dimensional evaluative structure distributes derivative incentives across heterogeneous domains:

$$\boldsymbol{U(a) = \sum_k \omega_k \times M_k(a)},$$

where metrics $\boldsymbol{M_k(a)}$ span structurally distinct evaluative dimensions.

By diversifying feedback channels, multi-metric oversight reduces proxy dominance and limits the capacity of improvement in one dimension to offset degradation in another. This reduces structural pressure toward threshold erosion.

10.4 Parameter Stabilization Mechanisms

Because drift operates at the level of parameter evolution, governance must address not only state trajectories but the dynamics of constraint sensitivity parameters $\boldsymbol{\theta_i(t)}$.

Under soft enforcement, sensitivity parameters evolve according to:

$$\boldsymbol{\dot{\theta}_i(t) < 0},$$

in expectation. Preventing misalignment therefore requires mechanisms that interrupt or counterbalance this monotonic decay.

Structural stabilization strategies may include periodic recalibration of constraint weights, independent enforcement layers insulated from optimization logic, automatic activation thresholds near boundary proximity, and architectural separation between performance maximization and constraint enforcement.

The unifying principle is the introduction of discontinuity into otherwise continuous derivative dynamics. By re-establishing invariant constraint manifolds, such mechanisms prevent endogenous weakening of boundary conditions.

Parameter stabilization functions as a higher-order control layer: it does not eliminate optimization, but constrains the long-run evolution of the optimization architecture itself.

10.5 Core Governance Insight

Alignment is an architectural problem of constraint design rather than a problem of intention.

Derivative-based systems, whether institutional or algorithmic, faithfully execute their objective structures. When constraints are compensable, delayed, or probabilistic, cumulative divergence is not an anomaly but the predictable long-run behavior of locally rational dynamics.

Sustainable alignment therefore requires constraint architectures that are non-compensable, temporally immediate, and structurally insulated from derivative trade-offs within the optimization objective.

The governance task is not to assume that locally coherent agents will remain globally aligned, but to design optimization environments in which boundary integrity remains dynamically invariant over time.

Alignment, in this sense, is a problem of stabilizing constraint manifolds within continuous derivative systems.

11. Conclusion

This paper introduced the Derivative-State Drift (DSD) framework as a unified structural account of cumulative misalignment in derivative-based optimization systems. The central claim is not that agents fail to optimize, but that optimization conducted over first-order state derivatives under soft constraint regimes generates endogenous drift as a matter of architectural necessity.

The analysis established that local derivative stability does not imply global normative alignment. A system may exhibit bounded state velocity, internal coherence, and continuous objective improvement while its normative distance increase monotonically. Divergence arises not from instability in state trajectories, but from gradual decay in constraint sensitivity parameters and the resulting deformation of the effective constraint manifold.

By instantiating the DSD framework in both elite institutional environments and artificial optimization systems, the paper identified a shared structural configuration: derivative-based evaluation rules, compensable constraint architectures, mediated or delayed feedback, and evolving threshold parameters. The comparison is architectural rather than anthropomorphic. Human and AI systems differ in substrate, cognition, and internal friction; yet when embedded within analogous optimization structures, they exhibit formally equivalent drift dynamics.

The governance implication follows directly from the model. Alignment cannot be secured through appeals to intention, moral character, or performance refinement alone. In derivative-driven systems, alignment depends on the stability of constraint architectures over time. When boundaries are soft, penalties compensable, and feedback incomplete, cumulative divergence is not accidental—it is structurally induced.

Derivative-State Drift is therefore not a pathology of particular actors, institutions, or technologies (Strogatz, 2018; Russell, 2019). It is a general property of optimization architectures in which local improvement is decoupled from invariant global reference conditions.

Misalignment, in this light, is not episodic failure. It is the predictable long-run behavior of locally rational systems operating within deformable constraint manifolds.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*. https://arxiv.org/abs/1606.06565

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*, 4299–4307.

Goodhart, C. A. E. (1975). Problems of monetary management: The U.K. experience. In *Papers in monetary economics* (Vol. 1, pp. 91–121). Reserve Bank of Australia.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*. https://arxiv.org/abs/1906.01820

Khalil, H. K. (2002). *Nonlinear systems* (3rd ed.). Prentice Hall.

Mahoney, J., & Thelen, K. (2010). *Explaining institutional change: Ambiguity, agency, and power*. Cambridge University Press.

Manheim, D., & Garrabrant, S. (2019). Categorizing variants of Goodhart's law. *arXiv preprint arXiv:1803.04585*. https://arxiv.org/abs/1803.04585

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review, 1*(6), 894–904. https://doi.org/10.2307/2084615

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.

Ostrom, E. (2005). *Understanding institutional diversity*. Princeton University Press.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering* (2nd ed.). Westview Press.

Thelen, K. (2004). *How institutions evolve: The political economy of skills in Germany, Britain, the United States, and Japan*. Cambridge University Press.

Weick, K. E. (1995). *Sensemaking in organizations*. Sage.

Williamson, O. E. (1985). *The economic institutions of capitalism*. Free Press.