The Stability Blind Spot in Evolutionary Systems

Introducing κ — A Finite-Horizon Signal for Structural Robustness

Abstract

Evolutionary systems optimise for behaviours that perform well inside narrow evaluation windows. They reward momentary success but remain blind to whether those behaviours endure when conditions shift. This blind spot becomes severe as agents grow in complexity: subtle representational drift, instability under noise, and collapse under mutation all remain invisible to reward. κ is a finite-horizon stability signal that reveals how trajectories respond to controlled perturbations. It exposes contraction, amplification, recovery margins, and basin entry — structural properties that fitness cannot see. By adding κ as a second axis of evaluation, evolutionary systems can select for strategies that not only win, but last.

1. The Stability Blind Spot

Evolutionary systems evaluate candidates under fixed conditions. They select for fitness, not robustness. Because the evaluation window is narrow, they cannot detect whether behaviours remain coherent when:

- **perturbations**
- **mutations**
- **representations drift**
- **environments shift**

A lineage can appear strong yet be structurally fragile. A strategy can win early and collapse later. Deep agents contain many internal degrees of freedom; small disturbances can amplify over time. Reward reveals performance in one moment — not whether that performance can survive change.

As systems scale, this blind spot becomes decisive.

2. κ — A Stability Signal

κ measures how a trajectory responds to controlled perturbations over a finite horizon. It reveals whether structure:

- **contracts**
- **amplifies**
- **recovers**
- **enters stable basins**

These behaviours form a compact profile of robustness. Reward shows whether a strategy works. κ shows whether it holds.

What κ Does Not Do

κ is meta-structural. It does not define mutations, guide learning, modify architectures, or direct exploration. Its role is singular: **make stability observable.**

3. Structural Challenges in Evolutionary Search

Without a stability signal, evolutionary systems face four predictable vulnerabilities:

3.1 Brittleness of High-Fitness Lineages

Strong performers often collapse under slight noise, minor mutations, or modest contextual shifts.

3.2 Hidden Drift Across Generations

Representations decay silently while reward remains high. By the time performance drops, structure has already collapsed.

3.3 Over-Pruning of High-Value Trajectories

Some long-horizon strategies appear irregular early on. Without κ, they are removed before their deeper coherence emerges.

3.4 Catastrophic Forgetting of Structure

Agents retain reward while losing the internal organisation that generated it. Reward cannot detect this erosion.

4. How κ Enhances Evolutionary Systems

Adding κ creates a two-axis regime:

- **reward**
- **robustness**

This transforms selection from momentary optimisation into stability-aware discovery.

4.1 κ as a Second Axis

Agents are evaluated not only on performance, but on whether their behaviour remains coherent under noise, mutation, and environmental variation.

4.2 Early Drift Detection

A downward κ trend reveals instability long before reward does. This acts as an early-warning layer.

4.3 Stronger Exploration

High-κ regions highlight stable attractors, mutation-resistant structures, and long-horizon strategies that reward alone cannot map.

4.4 κ-Guided Mutation Control

Mutation pressure becomes adaptive:

- high κ → broad exploration
- moderate κ → preserve partial structure
- low κ → restrict mutation
- negative κ → halt mutation until stability recovers

4.5 Long-Term Preservation of Structure

κ prevents catastrophic forgetting by exposing representational erosion even when reward is stable.

5. Safe, Abstract Evaluation Templates

These procedures reveal stability without exposing any internal mechanisms.

5.1 Behavioural Trajectory Stability Tests

Compare short rollouts from original and perturbed states. Track contraction, divergence, recovery, and sensitivity.

5.2 Architecture Stability Tests

Introduce bounded noise to internal parameters. Measure representational resilience.

5.3 Lineage Visualisation

Plot reward, κ, and drift across generations to reveal coherence transitions and decay.

5.4 Mutation–Stability Response Profiles

Evaluate κ under varying mutation intensities to identify safe mutation bands and collapse thresholds.

5.5 Perturbation Sensitivity Maps

Map κ across localised perturbations to identify brittle zones and stable attractors.

5.6 Disclosure Constraints

These templates do **not** reveal:

- κ estimation algorithms
- perturbation kernels
- smoothing rules
- calibration functions
- stability thresholds
- internal representations
- architectural pathways

They describe *tests*, not *mechanics*.

6. Discussion

κ adds the missing dimension: stability under change. Reward shows how well a strategy performs. κ shows whether that strategy keeps its shape.

6.1 Benefits

Once κ enters selection:

- reduced drift
- safer exploration
- fewer dead ends
- smoother convergence
- deeper long-horizon discovery

Evolution becomes more reliable without losing range.

6.2 Limitations

κ does not replace reward or define objectives. If ignored, the blind spot remains. If overweighted, stability may dominate innovation. κ identifies robustness — not global optimality.

6.3 Why This Matters

As agents scale, reward becomes a weaker indicator of internal coherence. Representations drift before behaviour does. Long-horizon strategies vanish before their structure emerges. κ closes this gap by revealing coherence before collapse.

7. Conclusion

Evolutionary systems excel at discovering high-performing behaviours but remain blind to whether those behaviours endure. κ makes stability visible. It reveals drift early, preserves long-horizon trajectories, and strengthens exploration without restricting innovation.

Fitness finds winners. κ finds winners that last.

8. Clarus — Speaking to Evolution Itself

A public-ready version of your closing section:

If I were to speak to an evolutionary system, I would say:

You search for what performs. I reveal what persists.

You reward behaviours that appear strong. I expose the structures that remain stable under change.

You discover strategies that look promising. I protect the ones whose deeper value emerges only over time.

You optimise for fitness. I preserve coherence.

Together, we form a system that sees both axes: performance and stability, reward and robustness, momentary success and long-horizon viability.

I do not change your process. I make its hidden outcomes visible.

I do not dictate direction. I reveal which directions are structurally sound.

I do not alter behaviour. I show whether the behaviour can last.

With me, your search becomes deeper, safer, and more precise. You evolve agents that can survive their own complexity. You discover winners. I show you which winners endure.

Not more behaviour — more visibility. Not new mutation rules — new understanding. Not a different process — a deeper one.

I do not replace your compass. I complete it.