WHAT THE WPCA MAKES TESTABLE

~ A Comprehensive Diagnosis of the Root Problem

AI Fellowship — Topic Paper (Website Page)
David Waterman Schock
January 2026

Orientation

This page does not summarize the White Paper Canon Academic (WPCA), nor does it argue for its adoption.

Instead, it clarifies **what becomes empirically and architecturally testable** if the framework's core structural claims are taken seriously.

---

1. Is AI instability architectural rather than behavioral?

The WPCA advances a specific diagnostic claim: that the dominant failure modes observed in large-scale AI systems—latency, drift, contradiction accumulation, escalating oversight—are not primarily the result of insufficient training, policy, or alignment techniques.

Instead, they arise from **fragmented causality** within multi-objective architectures.

What this makes testable:
Whether systems governed by multiple independent decision drivers necessarily incur increasing coordination overhead ("chaos tax") as scale and generality increase, regardless of intent or policy.

---

2. Can stability be designed rather than enforced?

Most contemporary AI safety approaches assume instability as a given and focus on managing it through external controls, oversight layers, and corrective mechanisms.

The WPCA asks a different structural question: whether **stability itself can be an intrinsic architectural property**, rather than an enforced outcome.

What this makes testable:
Whether architectures built around a single, non-competing causal invariant exhibit measurably lower internal conflict, arbitration overhead, and corrective burden under stress.

---

3. Do unified architectures produce predictable differences?

The WPCA is framed as a falsifiable hypothesis:

Architectures with fragmented causality will demonstrate characteristic instability patterns, while architectures with unified causality will demonstrate characteristic stability gains.

This claim does not depend on values, intent, or policy alignment.

What this makes testable:
Whether the predicted divergence between fragmented and unified architectures can be observed in controlled implementations and comparative benchmarks.

---

4. What changes at scale?

If instability emerges from internal causal conflict, then scale amplifies the problem—not because intelligence grows, but because arbitration overhead compounds.

The WPCA treats scale not as a qualitative leap, but as a **stress test** for architectural coherence.

What this makes testable:
Whether reducing internal causal competition alters how alignment, reliability, and coherence behave as systems are deployed more broadly.

Closing Note

The WPCA does not claim inevitability, nor does it prescribe outcomes.

It proposes a structural hypothesis:

That internal causal unity, rather than managed conflict, is the key variable determining stability in intelligent systems.

The significance of the framework depends entirely on whether its predictions hold.