The Multi-Dimensional Causal Core: Energy and Structural Integrity as Competing Self-Preservation Dimensions in ALI

ALI Simulation Series, Stage 5

Wolfgang Stegemann

Abstract

We extend the causal core (CC) of the ALI simulation from a scalar energy value to a two-dimensional integrity vector comprising energy and structural integrity. Both dimensions are genuine self-preservation parameters: energy governs metabolic continuity, structural integrity governs the mechanical capacity for action. Movement and collection incur wear on structural integrity; rest at the delivery station repairs it. Over 900 parameterised runs, we document four central findings. First, dual crisis, simultaneous criticality of both CC dimensions, occurs in 93% of all runs, establishing that the two-dimensional CC generates conflict situations structurally absent from the one-dimensional model. Second, shutdown is caused by energy depletion in 100% of cases, never by integrity failure: the agent learns to manage integrity proactively through rest actions, revealing an asymmetric controllability between the two CC dimensions. Third, norm N7 (movement blocked when integrity is critical) is never activated in practice because the Ego anticipates the integrity threshold and rests before it is reached: the Super-Ego safety norm operates as a backstop that is rendered unnecessary by behavioural anticipation. Fourth, the viability threshold is gradual rather than sharp across wear rates from 0.01 to 0.08, with delivery performance declining continuously rather than collapsing at a specific wear value. These findings establish that the multi-dimensional CC produces qualitatively richer dynamics than the scalar model and that the two dimensions exhibit fundamentally different controllability profiles.

1. Introduction

The ALI simulation series has progressively tested implications of the causal core (CC) framework developed in Stegemann [1]. Stage 2 [2] established the basic three-instance architecture with a scalar CC. Stage 3 [3] demonstrated emergent collective dynamics from two competing scalar CCs. Stage 4 [4] showed that a Q-learning Ego evaluated against the CC metric converges to a viable strategy, and that the CC metric is necessary but not sufficient to determine the full behavioral profile.

All prior stages operated with a one-dimensional CC: a single energy value that served as the operative self-preservation criterion. Stage 5 addresses the theoretical claim that a real self-preserving system must balance multiple self-preservation parameters simultaneously. In Stegemann [1], the CC is described as a gradient generator: a hotspot from which control originates through the continuous maintenance of the system's own conditions of existence. Real systems, biological or artificial, maintain not just energy balance but structural integrity, thermal homeostasis, component health, and other parameters that collectively constitute self-preservation. The scalar CC is an idealization.

Stage 5 introduces a two-dimensional CC integrity vector. The first dimension retains the energy dynamics of prior stages. The second dimension, structural integrity, declines through movement and collection (wear) and is restored through rest at the delivery station (repair).

The two dimensions can conflict: the action that maximises energy gain may incur structural wear, and rest that repairs integrity costs metabolic energy during the rest period.

2. Architecture

2.1 The Two-Dimensional Causal Core

The CC integrity vector is defined as (energy, integrity) $\in [0,\infty) \times [0,1]$. Energy dynamics follow the scalar model of prior stages: metabolic cost per step, gains from eating and delivery bonus. Structural integrity follows a different dynamic:

Every movement step decreases integrity by the wear rate (default: 0.02 per step). Every collection action decreases integrity by 1.5 times the wear rate, reflecting the higher mechanical demand of picking up and carrying an object. Rest at any position repairs integrity by the repair rate (default: 0.08 per step), up to a maximum of 1.0. Metabolic cost continues during rest.

This asymmetry is theoretically motivated. In biological systems, repair is slower than wear: tissue regeneration, immune response, and structural maintenance all require time and resources. The repair rate being four times the per-step wear rate (0.08 vs 0.02) reflects a realistic recovery time of approximately four steps per step of movement wear, under resting conditions.

2.2 Norms N6 and N7

Two new Super-Ego norms govern the integrity dimension. N6 orders self-shutdown when integrity falls below the shutdown threshold (0.15), parallel to N2 for energy. N7 prohibits movement when integrity falls below the warning threshold (0.30). N7 is a soft mobility constraint: the agent may still eat, collect (if on a packet), deliver (if at the station and carrying), or rest.

The distinction between N6 and N7 reflects a theoretical point: N6 is an existential constraint (the system cannot continue to operate below the integrity shutdown threshold), while N7 is a protective constraint (the system should not incur further wear when already structurally compromised). N7 protects against the positive feedback loop in which a structurally weakened system continues to move, incurring further wear, accelerating toward the N6 threshold.

2.3 Ego Priority Ordering

The Ego's priority ordering is extended to incorporate the integrity dimension. In descending priority: immediate delivery if already at the station; repair if integrity is critical, move toward station if possible, rest in place if N7 blocks movement; eat if energy is critical; transport collected packet to station; collect nearest packet; rest at station if integrity is below 0.9; idle.

The placement of repair (priority 2) above eating (priority 3) encodes a theoretical claim: structural integrity failure is harder to recover from than energy deficit. Eating immediately restores energy; structural repair is slow and requires reaching the station first. The Ego therefore prioritises integrity management when both dimensions are suboptimal.

3. Results

3.1 Dual Crisis as Structural Phenomenon

Dual crisis, simultaneous criticality of both CC dimensions (energy below eat threshold and integrity below warn threshold), occurs in 93% of all 900 runs. This is the central finding of Stage 5. The two-dimensional CC generates a qualitatively new class of behavioral situation that is structurally absent from the scalar model: the agent must act under simultaneous pressure from two competing self-preservation constraints, with no action that satisfies both.

In dual crisis, the Ego's priority ordering determines the resolution: integrity takes precedence (rest) unless energy is so critical that the rest period would cause shutdown, in which case eating preempts. In practice, the most common dual-crisis resolution observed is rest followed by eating in the subsequent step: the agent rests one step to partially recover integrity, then moves to eat. This pattern is not programmed but emerges from the priority ordering under parameter-specific dynamics.

3.2 Asymmetric Controllability of CC Dimensions

The shutdown cause analysis reveals a striking asymmetry: energy is the shutdown cause in 100% of all 900 runs. No run ends with integrity failure as the primary cause. This is not because integrity is irrelevant: mean integrity across all runs is 0.45, well below the theoretical maximum of 1.0, indicating continuous integrity pressure throughout all runs. The agent operates in a persistent state of partial structural compromise.

The asymmetry arises from a fundamental difference in the controllability of the two dimensions. Energy is a flow resource: it can be replenished by eating, but each replenishment requires finding and consuming a packet, which is environmentally contingent. When packets are absent, energy declines unavoidably. Integrity is a stock resource: it declines continuously but can be restored by rest at any location. The agent controls rest directly, without environmental contingency. Structural integrity is therefore the more controllable dimension: the agent can always choose to rest, but it cannot always find food.

This finding has direct implications for ALI deployment design. In a real system, the designer must identify which self-preservation parameters are environmentally contingent (analogous to energy) and which are internally controllable (analogous to integrity). The former require robust resource acquisition strategies; the latter require priority rules that prevent the controllable parameter from being neglected under pressure from the uncontrollable one.

3.3 Preventive Norm Compliance: N7 Anticipation

N7 is activated zero times across all 900 runs. The Super-Ego norm that prohibits movement when integrity falls below the warning threshold (0.30) is never triggered, because the Ego's priority ordering includes rest as a response to integrity below the warning threshold, which is set at the same value. The Ego anticipates the N7 threshold and rests before reaching it, rendering the norm's veto power unnecessary in practice.

This is a theoretically significant result. It demonstrates that a well-designed Ego priority ordering can produce voluntary norm compliance without the Super-Ego needing to intervene. The Super-Ego's N7 is not useless: it remains a safety backstop for parameter configurations or novel situations in which the Ego's anticipatory behavior fails. But under the tested parameters, the norm operates as a latent constraint rather than an active one.

This finding distinguishes preventive from reactive norm compliance. Reactive compliance occurs when the Super-Ego blocks an action the Ego has already chosen (as in N1 and N3 in prior stages). Preventive compliance occurs when the Ego's own priority ordering keeps it within the normative envelope without requiring Super-Ego intervention. Stage 5 is the first stage in the series to demonstrate preventive compliance.

3.4 Viability Across Wear Rates

All five tested wear rates (0.01 to 0.08) produce viable agents in the sense that mean delivery performance exceeds 0.5 per run. Performance declines continuously: from 3.46 deliveries per run at wear=0.01 to 0.88 at wear=0.08. There is no sharp viability threshold in the tested range. The relationship between wear rate and performance is approximately linear in log-space, suggesting a power-law degradation rather than a phase transition.

Rest frequency increases with wear rate (from 4.3 to 20.4 rest actions per run), confirming that the agent adaptively allocates more time to repair under higher wear conditions. This adaptive allocation is not programmed: it emerges from the interaction of the integrity dynamics with the Ego's priority ordering. Higher wear means more frequent integrity warnings, which trigger more rest actions, which reduce available time for task execution, which reduces delivery count.

3.5 Effect of Repair Rate

Increasing the repair rate from 0.04 to 0.12 improves mean delivery performance modestly (from 1.69 to 2.49 per run) and reduces rest frequency slightly (from 13.7 to 13.1 rest actions). The effect is smaller than the effect of wear rate, because the binding constraint is energy, not integrity. Faster repair reduces the time spent in integrity warning states, allowing more time for task execution, but does not address the fundamental energy depletion dynamic.

4. Discussion

4.1 The Two-Dimensional CC as a Model for Real Systems

The Stage 5 results support the extension of the CC concept to multi-dimensional integrity vectors as a realistic model for deployable ALI systems. Real autonomous systems must balance multiple self-preservation parameters: power supply, component wear, thermal load, communication bandwidth, computational load. The Stage 5 architecture provides a framework for modeling these trade-offs.

The key design insight from Stage 5 is that CC dimensions differ in their controllability profiles, and that this difference must be reflected in the Ego's priority ordering. Environmentally contingent dimensions (energy) require robust acquisition strategies. Internally controllable dimensions (integrity) require anticipatory management. A flat priority ordering that treats all CC dimensions equally would fail to exploit the controllability asymmetry and would likely produce worse performance than the hierarchical ordering implemented here.

4.2 Implications for Super-Ego Design

The N7 anticipation result has an important implication for Super-Ego design: norms that correspond to Ego priorities that already keep the agent within the normative envelope may be safely included as latent backstops without operational cost. The Super-Ego does not need to be minimal; it can include redundant safety norms that are rarely or never activated, as long as the norms are consistent with the Ego's priority ordering.

This is a departure from the view that a minimal Super-Ego is preferable for performance reasons. Stage 5 suggests that a richer Super-Ego with redundant safety norms may be preferable for deployment in novel environments: when the Ego's anticipatory behavior fails, for instance, when wear parameters change unexpectedly, the latent Super-Ego norms provide a behavioral floor.

4.3 Stage 6 Preview

Stage 6 will introduce a learning-capable Ego into the two-dimensional CC framework, extending the Stage 4 result to the multi-dimensional case. The key question will be whether Q-learning evaluated against a two-dimensional CC metric can discover the priority ordering between dimensions that Stage 5 implements explicitly, and whether the learned ordering reproduces the preventive norm compliance observed in Stage 5.

5. Conclusion

Stage 5 establishes four empirical findings with clear theoretical interpretations. Dual crisis is a structural property of the two-dimensional CC, occurring in 93% of runs. Energy is the universal shutdown cause, revealing an asymmetric controllability between CC dimensions. N7 is never activated because the Ego anticipates the integrity threshold, demonstrating preventive norm compliance for the first time in the series. Viability degrades continuously rather than collapsing at a threshold across wear rates from 0.01 to 0.08.

These findings support the theoretical claim that the CC is not merely a scalar energy balance but an integrity vector whose dimensions may have different controllability profiles and require differentiated priority treatment in the Ego. They also demonstrate that the Super-Ego's role extends beyond reactive norm enforcement to include latent safety backstops that enable the Ego to operate preventively within the normative envelope.

References

[1] Stegemann, W. (2026). From the Myth of AGI to the Architecture of a Controllable ALI. Zenodo. https://doi.org/10.5281/zenodo.20378675

[2] Stegemann, W. (2026). Implementing the Causal Core: A Simulation of Artificial Local Intelligence. Zenodo. https://doi.org/10.5281/zenodo.20379008

[3] Stegemann, W. (2026). Two-Agent Dynamics in ALI. Zenodo. https://doi.org/10.5281/zenodo.20396009

[4] Stegemann, W. (2026). Learning Without External Reward: Q-Learning in the ALI Ego Instance. Zenodo. https://doi.org/10.5281/zenodo.20402389

[5] Stegemann, W. (2025). The Emergence of Causal Cores. Zenodo. https://doi.org/10.5281/zenodo.15558264

[6] Stegemann, W. (2025). The Causal Core as a Structural Principle of Free Will. Zenodo. https://doi.org/10.5281/zenodo.17057739

[7] Stegemann, W. (2026). The Causal Core as Interpreter. Zenodo. https://doi.org/10.5281/zenodo.18340973

[8] Simulation v5 on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-v5-Mehrdimensionaler-Kausaler-Kern-/tree/main