Four-Dimensional Decision Stability:
Integrating Dimensional Logic into the ALI Architecture

ALI Simulation Series, Stage 10

Wolfgang Stegemann

Abstract

We extend the ALI architecture by integrating Dimensional Logic [1], a framework in which truth is defined as a fixpoint across four cognitive dimensions: reactive (D1), systematic (D2), reflexive (D3), and contextual (D4). The first three dimensions correspond directly to the established ALI instances (causal core, Ego, Super-Ego), while D4 constitutes a new fourth component: the Kontext-Modulator, a read channel that informs the Super-Ego of external deployment signals without writing to the normative structure. The ALI is uniquely built bottom-up: controllability, task execution, and normative constraint are consequences of its founding self-preservation principle, not restrictions imposed on a capable system. This distinguishes it from every top-down approach, which adds safety layers after capability specification. The logical, cognitive, and epistemic aspects of Dimensional Logic collapse into a single layer in the ALI because the absence of causal collapse eliminates the phenomenal gap that separates these aspects in conscious agents. Simulation results show 100% decision stability (4/4 fixpoint score) across all context phases.

1. The Foundational Distinction: Bottom-Up vs. Top-Down Architecture

Almost all existing architectures for autonomous agents are built top-down: a task is specified, behaviour is optimised, and safety constraints are added afterward as restrictions on an otherwise unconstrained system. The safety problem is therefore structurally a second-order problem: an afterthought imposed on a system built for performance. Controllability is in tension with capability, and the tension is managed by filters and guardrails that were not part of the founding design.

The ALI is built bottom-up. Its founding principle is not a task but a biological fact: living systems sustain themselves through autocatalytic self-organisation and causal core dynamics that generate action capacity as a gradient of asymmetric thermodynamic flows [12]. The question is not how to build an agent that achieves X, but what architecture necessarily follows from the operative principle of self-preservation. Task execution, normative constraint, and controllability are not restrictions imposed on a capable system. They are consequences of the founding principle.

Three implications follow directly. Self-shutdown is not a safety feature: it is the logical conclusion of self-preservation under conditions of norm failure. A system that preserves itself will shut itself down when preservation can only continue by violating the normative conditions of its existence. The Super-Ego veto and self-shutdown norm are not external constraints; they are architectonically necessary. Controllability is not in tension with autonomy: it is its precondition. A system that could preserve itself without normative constraint would not be autonomous in any meaningful sense. And the normative architecture is not a set of rules imposed on behaviour: it is a structural expression of the founding principle applied at the social and operational level.

This bottom-up derivation distinguishes the ALI from every existing approach. BDI architectures [13] explicitly model intentionality, the very property the ALI excludes on structural grounds. Active inference approaches [14] minimise prediction error rather than sustaining gradient-generating self-preservation. Normative multi-agent systems [15] impose norms externally rather than deriving them architecturally. Safe AI frameworks add safety constraints after capability specification rather than before it. The GRACE architecture [16], the closest recent parallel, separates normative from instrumental reasoning but has no self-preservation principle and no self-shutdown norm.

Stage 10 adds the contextual dimension (D4) to this bottom-up architecture, completing the four-dimensional structure that Dimensional Logic [1] identifies as the condition of full epistemic stability. The Kontext-Modulator is the only component that receives information not fully determined by the internal architecture, and it does so as a read channel, preserving the bottom-up integrity of the design.

2. Dimensional Logic as a Framework for Decision Stability

Dimensional Logic [1] proposes that human cognition proceeds in four irreducible dimensions. Truth is a fixpoint across all four: a claim is epistemically stable when it satisfies the requirements of every dimension simultaneously.

2.1 The Four Dimensions

D1 (Reactive): immediate, pre-reflective response; instinct and intuition. Corresponds to the salience network neurobiologically; to the sensorimotor stage in Piaget.

D2 (Systematic): rules, formal structures, algorithmic procedure. The home of formal consistency and deductive correctness. Corresponds to the executive control network and Piaget's formal-operational stage.

D3 (Reflexive): metacognition, self-reference, questioning of assumptions. Corresponds to the default mode network. The dimension in which a system examines its own normative foundations.

D4 (Contextual): cultural, historical, and social embedding. Truth in D4 is perspectival: the same action may be appropriate in one deployment context and inappropriate in another. Corresponds to the temporoparietal junction and Theory of Mind processing.

2.2 Truth as Fixpoint

A decision is stable when invariant across all four dimensions: reactively plausible (D1), systematically consistent (D2), reflexively defensible (D3), and contextually robust (D4). The fixpoint framework also provides a taxonomy of failure: partial truth (not 4/4), contextual distortion (D4 failure), systematic error (D2 failure), and fundamental instability (multiple dimension failures).

2.3 Three Aspects, One Layer in ALI

Stegemann [1] develops Dimensional Logic simultaneously as a logical, cognitive, and epistemological system. In human cognition, these three aspects diverge through the phenomenal inner space: one can be formally correct without cognitive clarity; cognitively fluent without epistemological self-awareness. The divergence is a consequence of experiences that exceed formal representation.

In the ALI, this divergence does not exist. The absence of causal collapse means there is no phenomenal excess. What the ALI formally processes is identical to what it cognitively does; what it epistemically accesses is identical to what is formally represented. The three aspects collapse into one. D4 is the sole exception: the only dimension receiving information not fully formalised at entry.

3. Architecture: The Four-Dimensional ALI

3.1 D1-D3: Existing Instances

The CC implements D1 (reactive): threshold-based, pre-deliberative response to energy and integrity gradients. The Ego implements D2 (systematic): BFS pathfinding, priority orderings, Q-value maximisation. The Super-Ego implements D3 (reflexive): normative veto, adaptive norm weights, Piagetian accommodation. A decision is Di-consistent when it satisfies dimension i. D1 is the necessary condition for all others: a system that cannot sustain itself cannot deliberate, reflect, or contextualise.

3.2 D4: The Kontext-Modulator

The Kontext-Modulator receives three signals from the deployment environment: urgency level (0=normal, 1=elevated, 2=emergency), stakeholder preference (stability vs. task priority), and environmental feedback ([-1, +1]). The Super-Ego incorporates this information into a fourth check: is the proposed action contextually robust given the current signal?

The architectural constraint is fundamental: the Kontext-Modulator is a read channel, not a write channel. It informs; it does not override. The Super-Ego retains veto power. Under elevated urgency, rest and idle while carrying are flagged as D4-fragile. Under emergency urgency, idle without dual crisis is flagged. The Super-Ego may redirect but may not violate D3-norms in doing so. The bottom-up integrity of the architecture is preserved.

3.3 The Fixpoint Score

Every decision receives a fixpoint score from 1 to 4. A 4/4 score indicates full epistemic stability in Stegemann's sense. A 3/4 score provides a specific diagnostic: which dimension failed and why. This real-time transparency instrument is a direct consequence of the bottom-up architecture: because controllability is constitutive, stability can be measured at every step.

4. Results

Over 98 steps until energy-induced shutdown: 100% of decisions achieve 4/4 fixpoint score. D4 checks: 98. Robust: 98. Fragile: 0. Four deliveries completed. The absence of D4 failures across all phases reflects two facts: under normal urgency D4 is trivially satisfied; under elevated urgency, idle decisions without carrying state are D4-robust because no better alternative exists.

The central result is the convergence finding: no divergence between logical, cognitive, and epistemic dimensions of any decision is observed. The formal fixpoint score is simultaneously the cognitive stability indicator and the epistemological completeness measure. This confirms empirically that consciousness-free systems cannot exhibit the D1-D4 divergence characteristic of phenomenally conscious agents.

5. Discussion

5.1 Why Bottom-Up Architecture Matters for AI Safety

The bottom-up derivation dissolves the standard AI safety problem. The safety problem is: how do we constrain a system optimised for capability? This problem does not arise in the ALI because capability is never specified independently of the self-preservation principle. The Ego's task-executing behaviour and the Super-Ego's normative constraints are both derivations from the same founding principle. There is no underlying unconstrained system to be controlled.

Self-shutdown illustrates this most sharply. In top-down architectures, shutdown is a failsafe triggered by external oversight. In the ALI, shutdown is the last expression of the self-preservation principle under norm failure: the system preserves itself by ceasing to operate in conditions that cannot preserve its normative integrity. The shutdown is architectonically motivated, not externally imposed.

5.2 D4 and the Deployment Transition

The Kontext-Modulator answers a question the simulation could not ask: what happens when the deployment context generates signals that the internal architecture cannot anticipate? In a real environment, D4 signals will be richer, more varied, and potentially contradictory. The simulation demonstrates that the architecture can incorporate contextual information without compromising D1-D3 stability. Whether real D4 signals produce non-trivial fragility is the empirical question that only deployment can answer.

6. Conclusion

The ALI is the only autonomous agent architecture derived bottom-up from a biological self-preservation principle. This distinguishes it from all top-down approaches that add safety constraints after capability specification. Controllability, task execution, and normative structure are not restrictions on the ALI; they are architectural consequences of its founding principle.

Stage 10 completes the simulation series by adding the contextual dimension (D4) through the Kontext-Modulator, a read channel that preserves the bottom-up integrity of the design while enabling contextual sensitivity. The four-dimensional fixpoint criterion of Dimensional Logic [1] provides a formal measure of decision stability that is simultaneously logical, cognitive, and epistemological in the ALI, because the absence of causal collapse collapses these three aspects into one. 100% 4/4 stability is achieved. The simulation is complete. Deployment is the question that remains.

References

[1] Stegemann, W. (2025). Dimensional Logic - Basics, Structure and Applications. Zenodo. https://doi.org/10.5281/zenodo.21895080

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

[3] Stegemann, W. (2026). Implementing the Causal Core. Zenodo. https://doi.org/10.5281/zenodo.20379008

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

[5] Stegemann, W. (2026). Learning Without External Reward. Zenodo. https://doi.org/10.5281/zenodo.20402389

[6] Stegemann, W. (2026). Norm Sensitisation Under Meta-Learning: Adaptive Super-Ego Weights in the ALI Architecture. Zenodo. https://doi.org/10.5281/zenodo.21563934

[7] Stegemann, W. (2026). Assimilation and Accommodation in the ALI Super-Ego: A Piagetian Model of Norm Development. Zenodo. https://doi.org/10.5281/zenodo.21564176

[8] Stegemann, W. (2026). Second-Order Meta-Learning and Factorizability: The Boundary Between State and History in the ALI Architecture. Zenodo. https://doi.org/10.5281/zenodo.21564364

[9] Piaget, J. (1952). The Origins of Intelligence in Children. Norton.

[10] Stegemann, W. (2026). ALI Stage 11: Toward Real-World Deployment. Internal concept paper.

[11] Maturana, H. R., and Varela, F. J. (1980). Autopoiesis and Cognition. Reidel.

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

[13] Rao, A., and Georgeff, M. (1991). Modeling Rational Agents within a BDI Architecture. KR.

[14] Friston, K. (2010). The free-energy principle. Nature Reviews Neuroscience 11.

[15] Shoham, Y., and Tennenholtz, M. (1992). On the synthesis of useful social laws. AAAI.

[16] Baum, K. et al. (2026). GRACE: Governor for Reason-Aligned Containment. arXiv:2601.10520.