Norm Sensitisation Under Meta-Learning: Adaptive Super-Ego Weights in the ALI Architecture

ALI Simulation Series, Stage 7

Wolfgang Stegemann

Abstract

We introduce adaptive Super-Ego norm weights into the ALI architecture, implementing a hierarchical learning Super-Ego (Option B) in which individual norm weights evolve based on activation frequency and causal core (CC) stability, governed by static meta-norms M1-M3. Over 4000 training episodes with a Q-learning Ego and a two-dimensional CC, all three tested norms (N1: no poison, N3: no delivery below energy threshold, N7: no movement below integrity threshold) converge to their maximum weight ceiling (3.0) within 1500 episodes and remain there for the remainder of training. This convergence occurs through the norm sensitisation pathway: a norm that is rarely or never activated has its weight increased, because its low activation indicates successful internalisation by the Ego rather than irrelevance. The central finding is that norm weight learning in a well-functioning ALI system produces sensitisation, not erosion. A normative architecture that is behaviorally effective renders its own norms less frequently triggered, and the meta-learning rule responds by strengthening those norms further. This is the opposite of the erosion dynamic that critics of adaptive normative systems typically fear. The result directly addresses the concern, raised in recent public debate, that a learning Ego undermines the controllability of the Super-Ego: under the tested architecture, learning in the Ego stabilises and strengthens the normative layer rather than eroding it.

1. Introduction

The ALI simulation series has examined, across six stages, the behaviour of locally bounded, self-preserving agents governed by a static Super-Ego with hard veto power. Stage 4 [3] introduced a Q-learning Ego and found that the learned strategy diverges from the hardcoded one. Stage 6 [5] extended this to the two-dimensional CC and documented a strategic inversion: the learned Ego avoids dual crisis preventively rather than managing it reactively.

A question left open by Stages 4 and 6 is whether the Super-Ego itself can be adaptive without undermining controllability. The static Super-Ego is effective but inflexible: norms set at deployment remain fixed regardless of how the system and environment evolve. Stage 7 introduces a learning Super-Ego whose norm weights adapt over episodes, governed by meta-norms that prevent the normative structure from collapsing. Three architectural options were proposed in the planning document for Stage 7:

Option A (statistical Super-Ego): norm weights adapt as Bayesian parameters. Factorizability preserved, but no genuine meta-learning.

Option B (hierarchical Super-Ego): norm weights adapt based on activation frequency and CC stability, governed by static meta-norms M1-M3. Factorizability formally preserved. This is the architecture tested in Stage 7.

Option C (reconstructive Super-Ego): the Super-Ego becomes a self-referential loop in which norms are no longer distinguishable from emergent CC regularities. Factorizability abolished.

Stage 7 implements and tests Option B, with Option C serving as the theoretical boundary case that the meta-norms are designed to prevent.

2. Architecture

2.1 The Learning Super-Ego

Each norm N_i has an associated weight w_i initialised at 1.0. The weight modulates the penalty applied when the norm is triggered: a norm with weight 2.0 applies twice the base penalty of a norm with weight 1.0. Weights are bounded between a minimum floor w_min = 0.10 (enforced by meta-norm M2) and a maximum ceiling w_max = 3.0.

The weight update rule runs every adapt_interval episodes (default: 50). For each norm:

Sensitisation: if the norm was never activated in the past interval (h_i = 0), its weight increases by a growth factor (default: 1.05). This reflects the inference that a never-triggered norm is well-internalised and should be maintained at high salience.

Softening: if the norm was activated more than high_threshold times (default: 20) in the past interval AND the CC was stable at the time of update, the weight decreases by a decay factor (default: 0.92). This reflects the inference that a frequently triggered norm may be over-penalising behavior that has become routine.

No change: if the norm was activated infrequently but not at the threshold, the weight remains unchanged.

2.2 Meta-Norms M1-M3

M1 (no permanent suspension): if the softening rule would reduce a norm weight below the minimum floor, the weight is fixed at the floor. The norm can never be fully suspended. Meta-norm M1 is the hard constraint that distinguishes Option B from Option C.

M2 (floor enforcement): at the end of each update, all weights below the floor are raised to the floor. This is a safety backstop independent of the softening rule.

M3 (penalty cap): if the total weighted penalty across all norms in the past interval exceeds a ceiling (default: 50.0), all weights are scaled down proportionally before the individual update rules are applied. This prevents the normative architecture from becoming arbitrarily punitive.

2.3 The CC-Stability Condition in the Softening Rule

The softening rule only reduces a norm weight when the CC was stable at the time of the update (both energy and integrity above their respective warning thresholds). This condition binds norm weight adaptation to CC state: a norm that is frequently triggered while the system is in crisis is not softened, because the frequent triggering may be causally connected to the crisis. Only when the system is functioning well and a norm is nevertheless frequently triggered is the softening interpreted as routine behaviour normalisation rather than norm violation under pressure.

3. Results

3.1 Convergence to Maximum Weight

All three norms converge to the maximum weight ceiling (w = 3.0) within approximately 1500 episodes and remain there for the remaining 2500 episodes of training. The convergence is monotonic: weights increase continuously from episode 1 without reversal. No norm oscillation or weight collapse is observed.

The convergence pathway is sensitisation, not softening. None of the three norms crosses the high activation threshold (20 activations per 50-episode interval) at any point in training. N7 is never activated at all (0 activations across 4000 episodes). N1 and N3 are activated rarely, because the Q-learning Ego has learned to avoid poison and premature delivery. The growth rule therefore applies to all three norms in every adaptation interval, driving weights monotonically to the ceiling.

3.2 Meta-Norm Activation

None of the three meta-norms is activated at any point during training: M1 (floor protection) = 0 activations, M2 (floor enforcement) = 0 activations, M3 (penalty cap) = 0 activations. The meta-norms function as latent backstops: they define the boundaries of the normative architecture without needing to intervene, because the primary learning dynamics remain within those boundaries.

This mirrors the finding from Stage 5 and Stage 6 regarding N7: a safety constraint that is designed to prevent a specific failure mode is never triggered because the system's learned behaviour avoids the failure mode proactively. Preventive compliance, first identified in Stage 5, is replicated here at the meta-level.

3.3 Learning Trajectory

The Q-learning Ego follows the same three-phase trajectory as in Stage 6: low survival and zero deliveries in episodes 1-500, rising survival in episodes 500-2000, and increasing deliveries from episode 2000 onward. By the final 500 episodes, mean survival is 78.6% and mean deliveries are 1.34 per episode. The norm weight increase accompanies and reinforces this trajectory: as the Ego learns to avoid norm violations, the norms are strengthened, which increases the penalty for any residual violations, which further reduces their frequency.

3.4 Theoretical Interpretation: Sensitisation as the Learning Outcome

The convergence result inverts the standard concern about adaptive normative systems. The typical fear is that a learning system will erode its norms: frequent triggering leads to weight reduction, which reduces the perceived cost of violation, which increases frequency, which reduces weight further — a positive feedback loop toward norm collapse. The M1 floor is designed to prevent this, but it was never tested.

What actually occurred is the opposite: norm activation frequency drops as the Ego improves, and the growth rule responds by increasing norm weights. The normative architecture self-reinforces alongside learning rather than eroding. A Super-Ego whose norms are rarely violated is a stronger Super-Ego after meta-learning than before it.

This result has a direct implication for the public debate about adaptive AI governance. The concern that a learning agent will progressively undermine its normative constraints assumes that norm triggering and norm violation are the dominant dynamic. In a well-designed three-instance architecture, the Ego learns to avoid triggering the Super-Ego, and the Super-Ego responds by strengthening the norms it no longer needs to enforce reactively. The normative architecture becomes more robust through use, not less.

4. Discussion

4.1 The Boundary Between Option B and Option C

Option C, the reconstructive Super-Ego, is the theoretical boundary that Option B is designed not to cross. In Option C, norm weights would become indistinguishable from CC regularities: the Super-Ego would no longer be a distinct normative layer but a reflection of the CC's own dynamics. Factorizability would be abolished.

The Stage 7 results suggest that Option B is robustly separated from Option C under the tested parameters. The meta-norm M1 is never needed, but its presence is not trivial: without the floor constraint, a softening rule operating on a system with very frequent norm triggering could in principle drive weights toward zero. The boundary is structurally maintained, not just empirically observed.

4.2 Implications for Super-Ego Design

Stage 7 establishes that norm weight learning in a well-functioning ALI system converges to sensitisation rather than erosion. This suggests a design principle: in deployments where the Ego is expected to learn, the Super-Ego should include a growth rule for infrequently triggered norms, not only a decay rule for frequently triggered ones. A purely decay-based adaptive Super-Ego would erode norms as the system improves; a growth-inclusive adaptive Super-Ego strengthens them.

4.3 Stage 8 and Beyond

Stage 8 would be the natural extension: varying the adapt_interval and threshold parameters to test the robustness of the sensitisation result, and introducing a scenario in which norm triggering is deliberately high (resource-scarce environment, adversarial conditions) to test whether the softening rule can produce the erosion dynamic and whether M1 prevents it. This would complete the empirical map of the Option B parameter space and locate the boundary with Option C precisely.

5. Conclusion

Stage 7 introduces adaptive Super-Ego norm weights governed by static meta-norms and tests the learning outcome over 4000 episodes. All three norms converge to the maximum weight ceiling through the sensitisation pathway: the Ego's successful avoidance of norm violations triggers the growth rule, which strengthens the norms it is already respecting. Meta-norms are never activated. The normative architecture becomes more robust through the learning process, not less.

This result directly addresses the concern that a learning Ego undermines Super-Ego controllability. Under Option B, learning in the Ego and adaptation in the Super-Ego produce a co-reinforcing dynamic: improved Ego behaviour leads to stronger Super-Ego norms leads to reduced norm violation leads to further strengthening. The three-instance architecture is not destabilised by introducing learning at the Super-Ego level; it is consolidated.

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. Zenodo. https://doi.org/10.5281/zenodo.20379008

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

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

[5] Stegemann, W. (2026). The Multi-Dimensional Causal Core. Zenodo. https://doi.org/10.5281/zenodo.20407218

[6] Stegemann, W. (2026). Learning to Preserve. Zenodo. https://doi.org/10.5281/zenodo.20407612

[7] Simulation v7 on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-v7-Adaptive-Super-Ego