Learning Without External Reward:
Q-Learning in the ALI Ego Instance

Evaluated Against the Causal Core Metric

ALI Simulation Series, Stage 4

Wolfgang Stegemann

Abstract

We replace the hardcoded BFS-based Ego of the ALI simulation (Stage 2) with a Q-learning component evaluated against an internal reward signal derived exclusively from the causal core (CC). No external reward is provided. The agent learns to act solely on the basis of what preserves the system. Over 3000 training episodes, we document three distinct learning phases: an initial survival-only phase, a transition phase in which task execution emerges alongside stable self-preservation, and a convergence phase in which deliveries increase but survival rate declines slightly. The central theoretical finding is that the CC metric as sole reward source is necessary but not sufficient to reproduce the explicit priority ordering of the hardcoded Ego. The learned Ego converges toward a slightly more aggressive task-execution strategy that sacrifices some self-preservation for higher delivery counts. This reveals that the architecture of the Ego contributes independently to the behavioral strategy: the explicit priority ordering in the hardcoded Ego (self-preservation before task) was not merely a convenience but a structural choice with behavioral consequences that Q-learning must approximate from reward signals alone.

1. Introduction

The ALI architecture developed in Stegemann [1] and implemented through simulation stages 1 to 3 [2, 3, 4] assigns specific roles to three instances: the causal core (CC) provides the operative self-preservation principle, the Ego mediates between CC and environment, and the Super-Ego encodes normative constraints. In Stages 1 through 3, the Ego was implemented as a hardcoded BFS pathfinding system with an explicit priority ordering: self-preservation takes precedence over task execution. This ordering was not learned but designed.

Stage 4 asks what happens when this ordering is not given. We replace the hardcoded Ego with a Q-learning component whose sole evaluation criterion is the CC metric: reward is derived entirely from changes in the agent's energy state and task delivery outcomes, both of which feed back into self-preservation. No external designer specifies which actions are good. The agent must discover the self-preservation-task balance from experience.

The central question is not whether Q-learning works, it does, but what it reveals about the ALI architecture. Specifically: does a learned Ego reproduce the behavioral profile of the hardcoded Ego, or does it converge to a different strategy? And what does the answer imply for the role of explicit priority ordering in Ego design?

2. Implementation

2.1 State Representation

The Q-learning Ego operates on a compact discrete state space with six components: energy level of the CC (four discrete levels: critical, low, medium, high), carrying state (binary), direction to the nearest energy packet (eight compass directions plus absent), direction to the delivery station (eight compass directions), distance to the nearest energy packet (five discrete levels: here, adjacent, near, medium, absent), and distance to the station when carrying (four levels). This yields a state space of approximately $4 \times 2 \times 9 \times 8 \times 5 \times 4 = 11{,}520$ possible states, of which 897 were visited during the 3000-episode training run.

2.2 Action Space

Eight actions are available: four directional moves (up, down, left, right), collect, eat, deliver, and idle. The distinction between collect (pick up for delivery) and eat (direct assimilation) follows the two-mode assimilation model introduced in Stage 2 [2].

2.3 Reward Function: CC-Derived, No External Signal

The reward function is derived exclusively from the CC state. No external designer defines which actions are valuable. The function has four components:

Delivery reward (+3.0 plus energy delta multiplier): completing a delivery to the station. The energy delta component reflects that delivery is more valuable when the agent is in good energetic condition.

Collection reward (+1.0 plus survival component): picking up an energy packet for delivery. This intermediate reward is theoretically justified: collection is a CC-relevant act because it initiates the action sequence that produces the delivery energy bonus. Without an intermediate reward, temporal credit assignment fails for the collect-transport-deliver sequence.

Assimilation reward (+2.0 plus energy delta): direct eating of an energy packet. Reflects the immediate CC benefit.

Survival-state modulator: +0.1 when energy is above the eat threshold (stable self-preservation), −1.0 when energy is below the eat threshold (critical state penalty). This modulator is applied to all non-assimilation rewards and encodes the CC principle directly: actions taken in critical energy states are penalized regardless of their intrinsic value.

Shutdown carries a reward of −8.0. Norm violations receive penalties of −1.0 (N1: poison collection) or −0.5 (N3: delivery when energy insufficient).

2.4 Learning Parameters

Learning rate $\alpha = 0.15$, discount factor $\gamma = 0.90$, initial exploration $\varepsilon = 1.0$ decaying by factor 0.995 per episode to a minimum of 0.05. Training ran for 3000 episodes of 200 steps each. The Q-table was updated online after each step.

3. Results

3.1 Three Learning Phases

The training trajectory divides clearly into three phases, each with a distinct behavioral profile.

Phase 1 (Episodes 1-500): Mean deliveries per episode: 0.01. Survival rate: 35%. The agent is in pure exploration; its actions are largely random. The low survival rate reflects the metabolic cost of random movement without a stable resource acquisition strategy. The small positive survival rate indicates the agent occasionally stumbles on a viable sequence.

Phase 2 (Episodes 501-1500): Mean deliveries: 0.10-0.48. Survival rate: 85-90%. A threshold is crossed: the agent has learned stable self-preservation (reflected in the sharp survival increase) and begins to discover the collect-deliver sequence. The survival stabilisation precedes the delivery increase, confirming that the CC metric is functioning as intended: the agent first learns what keeps it alive, then what fulfils its task.

Phase 3 (Episodes 1501-3000): Mean deliveries: 1.16-1.84. Survival rate: 66-85%. Deliveries increase continuously. Survival rate declines slightly. The agent is learning to execute deliveries more aggressively, accepting a somewhat higher risk of shutdown in exchange for higher task performance. The balance point is different from the hardcoded Ego.

3.2 Comparison with the Hardcoded Ego

At convergence (final 500 episodes), the Q-learning Ego achieves a mean of 1.84 deliveries per episode with a survival rate of approximately 67%. The hardcoded Ego of Stage 2 achieves approximately 5 deliveries per 200-step run with a survival rate that also terminates in shutdown, but later in the episode and after more deliveries. The Q-learning Ego is less efficient at task execution and depletes its energy faster.

The key structural difference is the priority ordering. The hardcoded Ego applies an explicit rule: if energy is below the eat threshold, suspend task execution and seek food. This rule creates a behavioral asymmetry, self-preservation always preempts the task, that the Q-learning Ego must approximate from reward signals. The survival-state modulator in the reward function encodes this asymmetry, but imperfectly: the modulator penalizes actions in critical energy states without prohibiting them, whereas the hardcoded priority ordering prohibits task-directed behavior in those states categorically.

The result is that the learned Ego adopts a mixed strategy: it sometimes continues task-directed behavior in mild energy deficit, accepting the penalty, when the Q-values for delivery are sufficiently high from prior experience. The hardcoded Ego never does this.

3.3 What the Q-Table Reveals

The Q-table converges to 897 visited states out of a theoretical maximum of approximately 11,520. This sparse coverage reflects the environmental structure: most states are not reachable under a coherent strategy, and the agent learns to operate in a limited region of the state space. The states that accumulate the highest Q-values for delivery actions are, predictably, those with carrying=1, station direction stable, and energy above the deliver threshold. The states that accumulate the highest Q-values for eat actions are those with energy level 0 (critical) and a nearby energy packet.

The Q-table thus encodes, implicitly, a version of the hardcoded priority ordering, but as a soft gradient rather than a hard rule. The learned Ego prefers survival when survival is most at risk, and prefers delivery when survival is stable. The difference from the hardcoded Ego is in the sharpness of the transition between these modes.

4. Discussion

4.1 The CC Metric as Internal Reward: Necessary but Not Sufficient

The central finding of Stage 4 is that the CC metric, used as the sole source of reward signal, successfully produces a learning trajectory that converges toward coordinated self-preservation and task execution. This confirms the theoretical claim of Stegemann [1]: a functional self-preservation criterion is a sufficient basis for purposive behavior in a locally bounded agent.

However, the CC metric alone does not fully determine the behavioral strategy. The explicit priority ordering of the hardcoded Ego, which is an architectural choice rather than a learned behavior, produces a sharper separation between self-preservation mode and task-execution mode. The Q-learning Ego approximates this separation but does not reproduce it. The architecture of the Ego thus contributes independently to the behavioral profile: the same CC metric, evaluated by a hardcoded Ego and a learned Ego, produces different strategies.

4.2 The Role of Explicit Priority Ordering

The hardcoded priority ordering in Stage 2 was introduced to resolve the norm-induced deadlock observed in Stage 1 [2]. It was a practical fix with theoretical grounding: self-preservation is the operative principle of the CC, and task execution is superordinate to it in the Super-Ego, not in the CC itself. The ordering encoded this asymmetry directly.

Stage 4 shows that this ordering carries behavioral weight beyond deadlock prevention. It shapes the entire strategy profile, not just the edge cases. A learned Ego without the explicit ordering converges to a viable but different strategy, one that is more aggressive in task execution and slightly less conservative in self-preservation. Whether this is better or worse depends on the deployment context. For safety-critical applications, the explicit ordering is preferable: it provides guaranteed behavioral bounds that a learned strategy cannot. For applications where task throughput matters more than individual survival, the learned strategy may be superior.

4.3 Implications for ALI Design

Stage 4 suggests a design principle for ALI systems that use learning components in the Ego: the reward function must encode not just the CC metric but also the structural priorities that the designer wants the Ego to respect. A survival-state modulator (penalising actions in critical energy states) is a step in this direction but does not fully replicate the categorical priority of the hardcoded Ego.

More generally, the results support the view that the three-instance architecture is not merely a conceptual framework but has concrete behavioral consequences. The assignment of priority ordering to the Ego, normative constraints to the Super-Ego, and the operative principle to the CC is not interchangeable. Moving the priority ordering into the reward function (and thus into the learning process) changes the behavioral profile. This has implications for the controllability of ALI systems: learned Egos are harder to bound behaviorally than hardcoded ones, and the Super-Ego's veto power becomes correspondingly more important as a safety backstop.

4.4 Stage 5 Preview

Stage 5 will extend the CC to a multi-dimensional integrity vector, introducing explicit conflicts between CC dimensions that the Ego must resolve. The Stage 4 finding, that the Ego's architecture shapes strategy independently of the CC metric, suggests that Stage 5 will need to specify how the Ego weighs competing CC dimensions. This weighting is itself an

architectural choice, not a learnable parameter from the CC alone, and will require explicit design decisions in the Super-Ego.

5. Conclusion

Stage 4 establishes that Q-learning in the ALI Ego instance, evaluated against an internal CC-derived reward signal, produces a viable learning trajectory with three distinct phases: survival learning, task emergence, and strategy convergence. The learned Ego achieves coordination between self-preservation and task execution without any external reward specification.

The central finding is that the CC metric as sole reward source is necessary but not sufficient to reproduce the explicit priority ordering of the hardcoded Ego. The learned Ego converges to a slightly more aggressive task-execution strategy. This reveals that the Ego's architecture, specifically whether self-preservation priority is encoded as a hard rule or approximated through reward shaping, contributes independently to the behavioral profile of the ALI system.

For controllable ALI deployments, this implies that learned Ego components should be combined with explicit architectural constraints rather than relying on reward shaping alone to encode safety-relevant priorities. The Super-Ego's normative architecture provides these constraints, but its effectiveness depends on the Ego respecting the structural priority of self-preservation over task execution, a priority that must be explicitly designed, not merely incentivised.

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. (2025). The Emergence of Causal Cores. Zenodo. https://doi.org/10.5281/zenodo.15558264

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

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

[7] Simulation v1 on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-Ein-minimalistischer-Prototyp-einer-Artificial-Local-Intelligence

[8] Simulation v2b on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-v2-Kausaler-Kern-als-operatives-Prinzip

[9] Simulation v3 on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-v3-Zwei-Agenten-Ressourcenkonkurrenz-und-N5

[10] Simulation v4 on GitHub: https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-v4-Lernfaehige-Ich-Instanz-Q-Learning/tree/main