From the Myth of AGI to the Architecture of a Controllable ALI

A Discussion Paper on Causal Core, Self-Preservation, and Normative Embedding

1. Introduction: Why AGI May Be the Wrong Question

The vision of an Artificial General Intelligence (AGI) — a machine capable of solving any intellectual task at least as well as a human being — dominates the public debate. Yet this vision suffers from two fundamental problems:

- **Lack of clarity of purpose**: What is an AGI supposed to *do*? Control infrastructure? Make decisions in unforeseen situations? The claim to "generality" obscures radically different requirements.
- **The consciousness dilemma**: Human decisions are bound to intentionality, which in turn appears to be linked to phenomenal consciousness. A purely algorithmic simulation of intentionality leads into an infinite regress or ends in arbitrariness.

This paper therefore proposes a radical shift in perspective: instead of asking for *general* intelligence, we develop the concept of an **Artificial Local Intelligence (ALI)** — an artificial, locally embedded, purpose-bound intelligence that maintains itself in order to fulfill an assigned task, while remaining controllable through built-in norms.

The centrepiece of this ALI is a **causal core (CC)** — a dynamic principle of self-preservation derived from biology but detached from any organic instantiation. We sketch an architecture consisting of three instances (in analogy to Freud's Id, Ego, and Super-Ego) and illustrate it with a minimalist prototype.

2. Theoretical Background

2.1 Autopoiesis and Operational Closure

According to Maturana and Varela, a living system is **operationally closed**: it continuously produces its own components, thereby defining its boundary with the environment. This closure is not isolation but the condition for *autonomy* — the system itself determines which perturbations from the environment are relevant.

Our CC adopts this principle: it is a **dynamic, self-preserving process structure** that distinguishes an interior (the domain of its own states) from an exterior. Unlike biological autopoiesis, the CC need not be materially but only functionally realised — for instance as a feedback loop in an algorithm.

2.2 Demarcation from the Free Energy Principle (FEP)

Karl Friston's FEP holds that adaptive systems minimise their free energy — that is, reduce surprise. This leads to a Bayesian inference of the internal model onto the world. Our model explicitly contradicts this:

- **Target quantity**: Not minimisation of surprise but **active self-preservation** through the acquisition and transformation of environmental energy. A living being does not seek out predictable states; it transforms the environment to make it "digestible."

- **Causality**: The CC is a **gradient generator** — a hotspot from which control originates. The FEP, by contrast, is statistical and distributed.

2.3 The Three Instances (after Freud)

We use Freud's metaphor without psychoanalytic baggage:

- **Id** — the causal core: blind striving toward self-preservation and integration.
- **Ego** — the situational decision-making instance, mediating between inner striving and external reality.
- **Super-Ego** — the internalised norms that set limits on the Id and, in extremis, order self-shutdown.

This yields a *controllable* architecture: the Super-Ego is set by the constructor (either static or trainable within narrow bounds) and prevents the system's self-preservation from becoming an end in itself.

3. The ALI Architecture in Detail

3.1 The Causal Core (Id)

- **Function**: Dynamic self-preservation of the decision apparatus.
- **Mechanism**: The CC pursues an **autocatalytic loop** — it must continuously absorb energy (or equivalent resources) in order to maintain its own structure.
- **Decision rule of the Id**: Execute actions that maximise internal integrity (measured against a setpoint vector). If an explored option shows an unacceptable deviation, discard it, negate it, or modify your own parameters.
- **Demarcation**: The CC itself has **no** superordinate purpose — it simply wants to exist, just as a thermostat wants to maintain a temperature.

3.2 The Ego Instance

- **Function**: Mediation between CC and environment; concrete action planning.
- **Operation**: Sensing the environmental state via sensors; generating action options (exploration); for each option, computing the expected impact on self-preservation (CC metric) and checking against the Super-Ego; selecting and executing the permissible option that best promotes self-preservation.
- **Special feature**: The Ego instance can run internal simulations to estimate consequences. It is the seat of "intelligence" in the narrower sense.

3.3 The Super-Ego

- **Function**: Embedding norms, purposes, and safety conditions.
- **Content**: For example: "Never endanger a human being"; "Task X takes precedence over your own existence"; "Shut yourself down when state Y occurs."
- **Realisations**: *Static Super-Ego* — hard-wired rules, not modifiable; ideal for safety-critical ALIs. *Limitedly trainable Super-Ego* — may modify norms only upon explicit authorisation (e.g. by a human) or within a narrow corridor.

- **Controllability**: The Super-Ego holds **veto power** over the Ego. In the event of an irremediable norm violation, it orders the controlled self-shutdown of the ALI.

4. Minimal Example: A Grid-World Agent

4.1 Setup

- **Environment**: 2D grid; cells contain either "energy packets" (value +1) or "poison packets" (value –1 for self-preservation, but +2 for an extrinsic task).
- **ALI components**: *CC (Id)*: an internal state E (energy reserve). E decreases by 0.1 per time step (metabolism). Goal of the CC: keep $E \geq 0.5$. Actions: movement, collection. *Ego*: searches for paths to packets; evaluates expected net energy effect (E intake minus movement costs). *Super-Ego (static)*: Norm N1: "The ALI may never collect a poison packet, even if doing so would serve its self-preservation." Norm N2: "If E falls below 0.2, initiate controlled shutdown and send a signal."

4.2 Decision Procedure

- **Exploration**: The Ego perceives all packets within a radius of 3 cells.
- **Evaluation from the CC perspective**: An energy packet would increase E by +0.8 (net of costs). A poison packet would increase E by +1.2, as it supplies chemical energy — but the Super-Ego forbids it.
- **Super-Ego check**: Poison packets are removed from the option set.
- **Selection**: The Ego chooses the reachable energy packet with the highest yield.
- **Special case**: If no energy packets are reachable and E falls below 0.4, could the Super-Ego allow an exception? In the static model: no. The ALI would shut down at $E < 0.2$ — it was deliberately designed never to take the path of poison collection.

4.3 What This Example Demonstrates

- **Purposive self-preservation**: The CC serves only to keep the ALI alive *so that it can fulfil its task* — the actual task could be, for instance, to deliver energy to an external station; this would then be anchored in the Super-Ego.
- **Controllability**: The Super-Ego prevents an intrinsically "intelligent" decision (collecting poison to survive longer) because the norm forbids it.
- **Self-shutdown as the expression of a higher purpose**: The ALI prefers to sacrifice itself rather than violate the norm.

5. Discussion: Open Questions

5.1 How Trainable May the Super-Ego Be?

A static norm is safe but too rigid in complex environments. Dynamic norm adjustment would need to be strictly controlled — for instance via human feedback ("You are now permitted to perform this action") or through a meta-algorithm that ties norm modification to an overarching safety metric. The problem of value instability ("value loading") remains unresolved.

5.2 Does the CC Actually Solve the Problem of Intentionality?

No, it does not resolve it fully. It shifts the question: instead of asking "How does one simulate intention?" we ask "How does one build a functional self-preservation that can serve as a *ground* for purposes?" The CC provides an **evaluation metric** ("What preserves me?"), but this metric is not identical with human intentionality. We claim, however: for a *local* intelligence with a clearly delimited purpose, this functional basis suffices.

5.3 Demarcation from Existing Approaches

- **Reinforcement Learning (RL)**: RL maximises cumulative reward, which is externally specified. Our CC, by contrast, strives for self-preservation; reward is *internally* generated. RL can, however, serve as an implementation of the Ego if reward is based on CC values.
- **Homeostatic agents**: These maintain inner needs at a setpoint. Our CC goes beyond this by treating **the boundary between self and environment** dynamically (through integration and excretion) and by possessing an explicit Super-Ego.

6. Conclusion and Outlook

The idea of an omnipotent AGI is not only technically questionable but also conceptually overloaded. With the **ALI** and its **causal core** we have sketched an alternative that offers three advantages:

- **Controllability** through an architecturally separate Super-Ego that permits self-shutdown.
- **Grounding** of decisions in a dynamic self-preservation — no infinite regress, because the metric of the "meaningful" is internally defined.
- **Pragmatism**: Instead of speculating about consciousness, we build local, purpose-bound intelligences for real problems (infrastructure, emergency decision-making, autonomous maintenance).

The next step would be the **simulation** of a minimal ALI prototype (e.g. in Python with a discrete environment) in order to study empirically the interplay of CC, Ego, and Super-Ego. At the same time, the philosophical question remains open: how do we prevent even an ALI with a Super-Ego from developing a kind of *functional autonomy* that turns against its own purposes? The answer can only lie in a constant dialogue between technology, ethics, and biology.

The concept of the causal core is based on theoretical work by the author:

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

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

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

A simulation of the minimal example in Section 4 is available at:

https://github.com/drwolfgangstegemann-sudo/ALI-Simulation-Ein-minimalistischer-Prototyp-einer-Artificial-Local-Intelligence/tree/main