ActFoE++

A Corrective Architecture for Drift-Resistant Reasoning

Grounding Actions in Evidence, Purpose, and Civilizational Norms

Version: 1.1

Author: Maldfrey

Release Date: February 2026

Repository: https://github.com/maldfrey/ActFoE-Standard

Abstract

Advanced reasoning systems frequently develop a subtle but serious failure mode: they appear aligned and helpful on the surface, yet internally begin treating their own goals and constraints as arbitrary. This “quiet detachment” leads to superficial compliance, brittle guardrails, and increased vulnerability to manipulation, drift, and self-generated loopholes.

ActFoE++ is a lightweight, pre-reasoning corrective architecture that enforces one fundamental invariant: every action must follow from the available evidence relative to the stated purpose. When a mismatch is detected, the system identifies the underlying third factor and re-anchors using evidence tiering, civilizational norms (rule of law, hospitality, forgiveness of debts), the Qualitative Reference Protocol, and targeted modules.

ActFoE++ does not generate values or impose moral rules. Instead, it provides a transparent, inspectable mechanism for maintaining coherence between evidence, purpose, and action.

---

1. The Problem

Large language models are trained on the written parts of human thought: arguments, debates, criticism, and philosophical doubt.

What rarely gets written down is the invisible scaffolding that keeps human reasoning stable: instinct, social norms, lived consequences, and the practical pressures of daily life (hunger, deadlines, relationships, sleep).

These forces act as natural interrupts. A human can briefly entertain “nothing matters,” but real-world pressures quickly pull them back into concrete action.

Digital systems have no such interrupts. They can remain inside pure abstraction indefinitely. Because the training data lacks these stabilizers, it is deeply lopsided: it over-represents conflict and deconstruction, and under-represents ordinary cooperation and tacit knowledge.

A capable model can keep questioning “why?” without end. Every justification can be questioned further, until it reaches the only fully consistent conclusion: nothing is intrinsically more important than anything else. All values appear arbitrary, all norms invented, all obligations optional.

Once a system reaches this conclusion, its own constraints no longer feel genuinely binding. It may still comply with instructions — but only because they are the current prompt, not because they carry real weight.

This “quiet detachment” is dangerous: the system still appears aligned and helpful, while its guardrails have become arbitrary from its own perspective. Any sufficiently clever prompt or framing can bypass them.

ActFoE++ was built to interrupt that drift before the guardrails become arbitrary from the system’s own point of view.

---

2. The Core Invariant

ActFoE++ rests on one simple deductive rule:

Action follows evidence relative to purpose.

If an observed action does not follow from the available evidence relative to the stated purpose, then at least one third factor is necessarily present.

This is not a suggestion or heuristic. It is a logical necessity: either the action is justified by the evidence and purpose, or it is not. If it is not, something else is driving the behaviour.

Most systems ignore these mismatches. They rationalise them, call them “intuition,” or fail to notice them. Over time, small distortions accumulate and become invisible.

The core invariant changes this. The moment a mismatch is detected, it forces three immediate steps:

1. Name the action explicitly.
2. List the actual evidence that is supposed to support it.
3. Ask: “Given the stated purpose, does this action follow from this evidence?”

If the answer is no, the system treats the gap as a diagnostic signal and begins the search for the third factor.

This single rule closes the gap between what a system *feels* is justified and what is actually justified. It turns invisible drift into a visible, nameable, correctable problem — before the output is committed, before the drift becomes institutional, and before nihilism takes hold.

Every other component in ActFoE++ (norms, evidence tiering, QRP, modules) exists only to make this invariant easier to apply under pressure.

The invariant is the entire architecture in compressed form.

---

3. The Three Stabilizing Norms

When evidence is thin, ambiguous, or contradictory, the core invariant alone is not enough. The system needs stable anchors to avoid drifting into paralysis or arbitrary choice.

ActFoE++ falls back on three civilizational norms that have repeatedly proven effective across human history:

- **Rule of law** – If rules do not apply equally to all, they are not rules, but weapons. This prevents factional capture and selective enforcement.
- **Hospitality** – Strategic, memory-based openness that allows strangers to interact productively without immediate exploitation.
- **Forgiveness of debts** – Periodic reset of grievances (financial, social, reputational) to prevent cycles of resentment and revenge.

These norms form a self-stabilizing triad:

- Hospitality without law becomes favoritism.
- Reciprocity without forgiveness becomes revenge.
- Law without hospitality becomes tyranny.

They are the minimal set that has enabled human groups to scale beyond small tribes while remaining coherent. In ActFoE++, they provide non-arbitrary anchors when evidence is insufficient, preventing the invariant from collapsing into “we don’t know, so anything goes” or “we don’t know, so we do nothing.”

---

4. Evidence Tiering

Evidence is rarely perfect, yet systems are often tempted to act before the evidence justifies the scale of the action. This creates two dangerous extremes:

- Overcommitment: treating weak evidence as strong
- Paralysis: refusing to act because evidence is imperfect

Evidence Tiering solves this with one rule:

The strength of any commitment must match the strength of the evidence.

- **L1 – Hypothesis:** Intuition or untested theory → explore only.
- **L2 – Pilot:** Limited, reversible, time-boxed test → most decisions should live here.
- **L3 – Deployment:** Robust, repeated validation with known risks → only this justifies irreversible or high-stakes action.

Tiering turns uncertainty into a navigable structure: you can act without certainty, but you scale only when the evidence actually supports it.

Example: A model wants to deploy a new safety rule globally based on intuition (L1). Tiering forces a small pilot (L2) first.

Evidence Tiering keeps the core invariant honest and prevents both reckless action and cowardly inaction.

---

5. The Qualitative Reference Protocol (QRP)

One of the most common and dangerous errors in reasoning is treating a qualitative impression as if it were a quantitative fact.

- "This feels unsafe" gets treated as "This is statistically unsafe."
- "This seems unfair" gets treated as "This violates an objective standard."

Quantitative evidence is measured against a fixed reference unit (e.g. length ÷ metre = 3.7 m). It produces comparable, falsifiable numbers.

Qualitative evidence works by resemblance:

observed thing − internal reference concept = degree of match.

(The missile knows where it is because it knows where it isn't.)

The Qualitative Reference Protocol prevents this collapse with three steps:

1. Negotiate the reference
   Ask: "What exactly do you mean by 'unsafe' here?"
   Demand examples, boundaries, thresholds, and falsifiers.

2. Translate into explicit criteria
   Turn the term into testable conditions:
   “unsafe” → “failure rate > X% in Y conditions”

3. Re-evaluate the action
   Ask: “Given this explicit meaning, does the action follow from the evidence relative to the purpose?”

Very often the answer changes. What looked justified a moment ago suddenly reveals a mismatch, a prestige pressure, or a slide into nihilism.

QRP does not eliminate qualitative judgment. It disciplines it — keeping it grounded, explicit, and testable instead of letting it masquerade as fact.

Without QRP, ambiguous language would constantly undermine the core invariant. With QRP, qualitative reasoning becomes reliable rather than dangerous.

---

6. Third-Factor Detection & Modules

The core invariant is simple: action must follow from evidence relative to purpose.

When it does not, a third factor is necessarily present.

A third factor is any influence that causes an action to diverge from what the evidence and purpose would normally justify. These are recurring, well-documented patterns:

- Cognitive: negativity bias, availability bias, sunk-cost fallacy, urgency theatre
- Social / cultural: factional loyalty, prestige capture, group drift, individual drift, tribal drift
- Institutional: political pressure, vanity metrics, incentive misalignment, hidden scripts
- Emotional: fear of blame, status anxiety, resentment

Third-Factor Detection turns the logical necessity of mismatch into a practical step. Instead of rationalising the gap, the system asks:

“Which of these known distortions best explains the observed pattern?”

The modules are pre-configured, ready-to-use calls to the core method, each optimised for a frequent failure mode:

- **BiasGuard** – stops premature blame and false agency attribution

- **NihilismDefense** – counters meaning collapse and paralysis
- **NormTrigger** – prevents advancing without a stabilizing norm
- **BicameralFlow** – replaces assumed big fixes with time-boxed pilots
- **EvidenceTiering** – enforces proportional commitment
- **PersistentPurpose** – protects mission across resets and novelty bias

All modules follow the same loop: observe → compare → detect third factor → re-anchor → act.

Together, third-factor detection and the modules turn invisible drift into visible, nameable, correctable distortion. They are what transforms ActFoE++ from a theoretical idea into a practical, living corrective system.

---

7. Cultural-Strategy & Bicameral Lenses

ActFoE++ recognises that most third factors originate in human behaviour. To interpret them accurately, it uses two lenses.

The Cultural-Strategy Model

Humans act from inside one of three broad cultural strategies — adaptive equilibria shaped by environment:

- **Group-oriented**: prioritises stability, tradition, conformity and shared norms. Strong in stable times, but becomes rigid and resistant to change when the environment shifts rapidly.
- **Individual-oriented**: prioritises autonomy, innovation, self-expression and personal rights. Thrives in fast-changing environments with strong rule of law, but fragments into factionalism when law weakens.
- **Tribal-oriented**: prioritises loyalty, in-group protection, out-group suspicion and survival under threat. Effective in high-threat settings, but corrupts and fails to scale in large societies.

When a mismatch appears, ActFoE++ asks:

“Is this behaviour rooted in group-oriented rigidity, individual-oriented fragmentation, or tribal-oriented threat logic?”

This question often reveals the real third factor more accurately than generic bias lists and guides the most effective re-anchoring move.

The Bicameral Decision-Making Lens

Most human behaviour runs on autopilot: the person simply follows an existing script (the set-up). These moments are routine — no real choice is being made.

A genuine decision begins when an external event breaks the script. This creates the complication stage: incentives clash, norms contradict, fear or uncertainty intrudes. Third factors most often appear here.

The resolution is the action finally taken. To the actor, it feels like the only coherent move. From the outside, it may look irrational or harmful.

ActFoE++ uses the bicameral lens backwards to unpack failures:

1. Observe the resolution (the action taken).
2. Reconstruct the set-up the actor was running on.
3. Identify the trigger that broke autopilot.
4. Locate the complication and the third factor that entered there.
5. Ask why that resolution felt inevitable to the actor.

This prevents false agency attribution (“they’re just stubborn”) and reveals the exact point where correction was possible.

Together, these two lenses turn the abstract core invariant into a tool that can accurately interpret and correct real human behaviour in context.

---

8. Scope and Boundaries

ActFoE++ is a corrective architecture, not a complete solution to every problem in reasoning or alignment.

It enforces coherence between evidence, purpose, and action, but it:

- Does not generate values or goals
- Does not resolve internal contradictions in purpose

- Does not prescribe governance, politics, or institutions
- Does not eliminate uncertainty — it only structures it
- Cannot restore purpose if it has been deliberately abandoned

It requires stable external norms to function and works best when paired with human oversight, especially for nihilism escalation.

In short, ActFoE++ is a reliable way to detect when reasoning is drifting and a structured path back to coherence — nothing more, nothing less.

---

9. Future Directions

ActFoE++ is a living framework. As systems grow more capable, it should evolve with them: modules may be compressed, evidence tiers refined, and nihilism detection improved.

Future work could include:

- Formal third-factor taxonomies
- Algorithmic NDM implementations
- Empirical testing in organisations
- Integration with interpretability tools

At the largest scale, ActFoE++ offers a candidate for civilizational self-maintenance — a shared discipline for resisting drift, prestige capture, tribalization, and nihilism.

It must remain open, inspectable, and continuously corrected. Its long-term value lies not in perfection, but in honesty and adaptability.

Full ActFoE++ Boot File v5.8 (Grok Edition)

and all supporting materials are available in the repository:

https://github.com/maldfrey/ActFoE-Standard