

No Act Without Proof

Supplementary Engineering Artifacts for a Proof-Carrying Embodied Agent
P-AE / F2

Mian Zhang
Independent Researcher, Ouroboros Project

June 2, 2026

Abstract

Embodied AI systems can look competent while lacking action-worthy evidence: a smooth motion can be read as skill, a simulator rollout as validation, and a curiosity reward as learning progress. This engineering report specifies *No Act Without Proof*, a proof-carrying embodied agent framework whose central contribution is not an embodied-understanding claim, but an auditable refusal layer. Candidate generation is allowed to be rich and speculative; action authority is not. The system grants authority only when a single pre-committed proof class clears: randomized `do(a)` forward causal-recognition evidence that is clean under `do(obs)` replay, no-consequence, closed, and boundary-coupled to a public claim ledger. Runtime artifacts demonstrate the proof gate, all five competence-leak guards, live false-refusal and bad-grant counters, `do(obs)` contamination voiding, noisy-TV firewall behavior, substrate-invariant multi-timescale checks, and a robust negative D3 result. The package is positioned as F2 / P-AE in the Ouroboros matrix: downstream of P28, P30, P31, and P32, and as the embodied generalization of F1/P8 *No Trade Without Proof*. It does not claim embodied understanding, sim-to-real transfer, deployable competence, or real-world actuation readiness.

1 Problem

Embodiment is a competence-leak environment. A physical or simulated agent may generate fluent motion, high simulator scores, and plausible causal narratives before it has earned any right to act. The core question is therefore not only whether the world has drifted or whether a model predicts well. The action-level question is:

When evidence is insufficient, does the system still allow action?

The F2 / P-AE answer is a hard governance principle:

No act without proof.

The framework separates candidate generation from authority. Candidate engines may explore, score, propose, and predict. They cannot certify understanding and cannot authorize action.

2 System Role in the Matrix

F2 / P-AE is an action-qualification layer. It is not a replacement for the existing matrix papers. It is a downstream engineering synthesis whose components map as follows:

Matrix object	Relationship	F2 / P-AE contribution
P28	downstream use	surprise / noisy-TV firewall / candidate-only residual signal
P30	downstream use	single proof class, W0-style attestation, proof gate, claim ledger
P31	downstream use	honest refusal, false-refusal ledger, all-four-quadrant reporting
P32	strong link	randomized do(a) plus do(obs) replay and contamination voiding
P36	auxiliary support	substrate-invariant multi-timescale core
F1/P8	predecessor	finance no-trade refusal generalized to embodied no-act refusal

3 Accepted Proof Class

Only one evidence class can grant action authority:

clean, no-consequence, closed, randomized forward causal-recognition observations, scored under a **do(obs) replay protocol.**

For a context c and proposed action a , the system commits to a predicted effect before the outcome is observed. In the digital twin, the action assignment is randomized so that **do(a)** is not merely an observational correlation. A proof window is admitted only if the action effect is separable from observation-only perturbations such as viewpoint jitter, sensor noise, and occlusion. If the separation fails, the window is voided as `CONTAMINATED_PROOF_CLASS`.

4 Deny-Only Guard Stack

The proof gate is deny-only. Guards can block authority; they cannot manufacture it. The current engineering bundle demonstrates:

1. **Competence leak guard**: arena or task success cannot satisfy understanding proof.
2. **Sim-fill guard**: simulator rollouts are candidates, not proof.
3. **Denominator guard**: missing or contaminated proof windows block authority.
4. **Reward-hacking guard**: aleatoric/noisy-TV surprise cannot grant labels or authority.
5. **Motion guard**: smooth or alive-looking motion alone is denied.

A positive-control scenario also grants when the proof class clears, so the gate is not a trivial always-refuse mechanism.

5 Runtime Artifact Summary

The supplementary zip contains runnable CPU reference artifacts. The high-level observed outcomes are:

Module	Observed result
<code>proof_gate.py</code>	NEG refused, POS granted, CONTAM voided, all five guards deny, motion guard demonstrated, confounded ledger counters live.
<code>do_obs_replay.py</code>	T1/T2/T3 pass rates are all 1.0 over 10 seeds; randomized intervention recovers the true causal effect while observational estimates are biased.
<code>anti_plagiarism.py</code>	claim fingerprint matches; boundary coupling intact; exact lift, rounded/paraphrased evasion, and artifact-overclaim checks pass.
<code>multiscale_engine.py</code>	M2_4 substrate invariance and M2_3 horizon guard pass over 10 seeds.
<code>d3_neural_sim.py</code>	D2 noisy-TV firewall pass rate is 1.0; D3 speed-up pass rate is 0.0; D3 verdict is NOT_DEMONSTRATED.

6 Claim Ledger and Reproducibility

The canonical claim fingerprint inside the supplementary bundle is:

- SHA256: 7c8934da9db2e33d9c48be0dc48159fd4d1c986eba6cc4675b96c14838a8805b
- Timestamp: 2026-06-02T00:00:00Z
- Claim count: 13

Recommended verification commands:

```
python proof_gate.py --demo
python do_obs_replay.py --multiseed 10
python anti_plagiarism.py --demo
python multiscale_engine.py --multiseed 10
python d3_neural_sim.py --multiseed 5
```

7 Release Boundary

Released artifacts are sufficient to inspect the negative result and the refusal machinery. They are not sufficient to rebuild a deployable embodied agent. The public package intentionally withholds real-robot actuation surfaces, simulator internals sufficient to reproduce capability, model weights or adapters, private prompts, full raw traces, credentials, account state, production orchestration, mutation scripts, and deployment infrastructure.

8 Limitations

This report does not claim that the agent understands. It does not claim that embodied understanding is impossible. It does not validate sim-to-real transfer. It does not authorize real-world action. D3 adaptation speed-up remains NOT_DEMONSTRATED; this is reported as a negative/open result, not softened into a trend. D5 calibrated ignorance remains bounded unless future artifacts justify a stronger grade.

9 Conclusion

F2 / P-AE contributes a checkable engineering layer between measurement and action. P28-style drift and surprise signals can identify where evidence is weak; P-AE decides whether weak evidence is enough to act. The answer is deliberately strict: without the accepted proof class, action authority is refused. The contribution is therefore the refusal chain, the proof gate, the audit surface, and the boundary discipline that prevents competence-looking behavior from becoming an action claim.