A■ vs. Free Energy Principle (FEP) ## A strict scientific positioning, proof sketch, and implications for building AI

Summary (what this document does) This document (1) states a **precise mathematical relationship between the A■ invariant and the Free Energy Principle (FEP), (2) explains why treating FEP as *fundamental* for "intelligent AI" is methodologically fragile, and (3) shows how **A■ closes the gap** by re-framing prediction as an optional, costly operator under admissibility.**

Key stance: this text does **not** dispute FEP as a powerful descriptive framework. It disputes **FEP as a fundamental primitive**.

---

1. Preliminaries: the A■ invariant (discrete, admissibility-first)

Let $S_t\in\mathcal S$ be the state of an open system at time t. Let $\mathcal A(S_t)$ be the locally available transitions ("actions"). Let
- $H_t(S_t,a)\ge 0$: instability / uncertainty / drift proxy,
- $T_t(S_t,a)\ge 0$: transition risk / irreversibility proxy,
- $Z_t(S_t,a)\ge 0$: execution impedance proxy.

1.1 Admissibility Fix thresholds $H_{\mathrm{crit}},T_{\mathrm{crit}}$. Define the admissible set: $$\mathcal A'_t(S_t)=\{a\in\mathcal A(S_t)\mid H_t(S_t,a)\le H_{\mathrm{crit}}\ \wedge\ T_t(S_t,a)\le T_{\mathrm{crit}}\}.$$

1.2 Local relaxation (A■ rule) If $\mathcal A'_t(S_t)\neq\emptyset$, the realized transition satisfies: $$a_t^*\in \arg\min_{a\in\mathcal A'_t(S_t)} Z_t(S_t,a).$$ If $\mathcal A'_t(S_t)=\emptyset$, the system collapses to a termination / quiescent state (a valid fixed point).

Interpretation: $\arg\min$ denotes a fixed point of local relaxation, not deliberation or global planning.

---

2. FEP: variational free energy as an instrumental bound

Let o_t be observations and s_t latent states. Let $p_\theta(o_t,s_t)$ be a generative model and $q_\phi(s_t)$ a variational posterior. The variational free energy is typically defined as:
$$
F(q_\phi,\theta) \;=\; \mathbb E_{q_\phi(s)}[-\log p_\theta(o,s)] \;+\; \mathbb E_{q_\phi(s)}[\log q_\phi(s)] \;=\; D_{\mathrm{KL}}(q_\phi(s)\,\|\,p_\theta(s\mid o)) \;-\; \log p_\theta(o).
$$
Minimizing F in q_ϕ tightens an upper bound on surprisal $-\log p_\theta(o)$.

In many FEP presentations, *action* is selected to make future observations less surprising under the generative model ("active inference").

---

3. The core logical issue: why “FEP as fundamental” is fragile for AI

Treating FEP as fundamental for building intelligent AI typically introduces hidden primitives that A■ avoids:

3.1 FEP requires a privileged internal model class FEP is defined relative to a generative model \(p_\theta\). In AI engineering, this implies: - a privileged model class, - a belief-update mechanism tied to that class, - a notion of “good action” derived from the model.

If \(p_\theta\) is misspecified (inevitable in open worlds), minimizing \(F\) can produce **confident but wrong stabilization**: the system reduces its bound, not external instability.

3.2 FEP encourages teleology if used as a design primitive When implemented operationally, “minimize free energy” is often encoded as goal-like optimization over policies. This can re-introduce: - implicit agency, - long-horizon planning as a primitive, - reward-like structure (even if renamed).

This is not a flaw of FEP as a descriptive theory; it is a **methodological hazard** when used as a foundational design recipe for AI.

3.3 FEP does not natively give a *hard admissibility gate* Real deployments need admissibility (safety/irreversibility/uncertainty gating) **before optimization. FEP, as commonly operationalized, tends to “optimize through” uncertainty (by adjusting beliefs/policies) rather than declaring transitions non-realizable.**

3.4 FEP does not natively make “non-generation / silence” a first-class stable outcome LLM failure modes (hallucination, overgeneration) are strongly driven by “must answer” pressure. A system that treats prediction/answering as mandatory will fabricate. FEP-style action selection does not by itself define a universal, mandatory collapse-to-silence fixed point when admissibility is indeterminate.

---

4. A■ closes the gap: prediction becomes an optional, costly operator

A■ reframes prediction as an operator \(F\in\mathcal A(S)\) with its own cost components:
- \(Z(F)\): compute/latency/tokens,
- \(H(F)\): model uncertainty/variance,
- \(T(F)\): risk of triggering unsafe downstream transitions.

4.1 Prediction admissibility Prediction is admissible only if it satisfies the same admissibility gate: \[H(F)\le H_{\mathrm{crit}},\quad T(F)\le T_{\mathrm{crit}}, \] and it reduces near-term instability (operationally measured in the next-step neighborhood), otherwise it is not invoked.

4.2 The design implication for AI - **FEP can be retained as *one possible internal mechanism* for the prediction operator \(F\) (a way to compute a stable estimate). - **A■** remains the fundamental invariant: it decides whether prediction is admissible at all, and whether**

silence/termination is the correct stable remainder.

This prevents “optimize through uncertainty” and supports principled abstention.

---

5. Formal relationship: FEP as a realization inside the A■ envelope

Proposition 1 (FEP-minimization is A■-consistent under admissibility) Assume a system has a prediction operator \(F\) that updates \(q_\phi\) to reduce \(F(q_\phi,\theta)\) (variational free energy) and assume: 1. the prediction operator is invoked only when admissible: \(H(F)\le H_{\mathrm{crit}}, T(F)\le T_{\mathrm{crit}}\); 2. local execution impedance \(Z\) is minimized among admissible actions; 3. the instability potential \(\Phi\) is monotone aligned with admissible \(Z\)-ordering (Lyapunov/supermartingale condition as in the A■ framework).

Then the combined system trajectory is A■-consistent, and FEP appears as an internal mechanism for computing admissible stabilization steps, not as a fundamental primitive.

Proof sketch. Under (1), \(F\) is a member of \(\mathcal A'(S)\) only when admissible. Under (2), the realized action minimizes execution impedance among admissible options. Under (3), the chosen transitions yield non-increasing expected instability potential, hence follow the A■ relaxation rule. The variational free energy minimization is internal to \(F\) and does not alter the outer admissibility-first ordering. \(\square\)

Corollary (A■ is strictly more general than FEP) A■ does not require any generative model \(p_\theta\) or variational family \(q_\phi\). Therefore A■ applies to systems where prediction is absent, disabled, or physically meaningless, whereas FEP does not.

---

6. What should be clarified in your statement (small but important)

Your formulation is essentially correct. To be maximally defensible, add two clarifications:

1. **“FEP is not fundamental” means:** prediction/inference is *not necessary* for stabilization; it is one possible mechanism.
2. **A■ is a stability invariant, not a truth principle:** A■ guarantees elimination of unstable trajectories, not semantic correctness. Correctness may emerge as a byproduct when the environment and constraints make false constructs unstable.

---

7. Practical consequence for LLMs - LLMs already perform local next-token relaxation in a learned landscape. - Failures occur when the system is pressured to continue without admissible stabilization (high \(H\), high \(T\)). - A■-first design makes termination a stable fixed point and treats prediction/explanation as optional.

---

8. Conclusion FEP remains valuable as an internal computational tool (a way to generate stable estimates). However, using FEP as a *fundamental design primitive* for AI risks reintroducing teleology, implicit agency, and optimization-through-uncertainty failure modes.

A■ provides a strictly more general invariant-level envelope: admissibility precedes optimization; prediction is optional and costly; silence is a valid stable remainder when no admissible continuation exists.