Bounded Fallibility

A structural argument for deterministic admissibility boundaries in AI execution systems

---

Core claim. In sufficiently complex, open-world deployments, residual error at novel inputs is inevitable. The architectural response is not better confidence scores, but an independent deterministic admissibility boundary that decides whether an AI proposal may proceed to execution.

Scope. This note is deliberately narrow: it covers the execution layer (pre-action gating), not model training, not alignment optimization, and not normative governance of values. It targets systems where incorrect actions carry material consequence: engineering, infrastructure, finance, regulated operations.

Vocabulary

`Proposal`
: A candidate action, design, or decision produced by an AI system.

`Admissibility`
: Whether a proposal satisfies a set of formal constraints (physics, code, limits).

`Fail-closed`
: Uncertainty or under-specification yields rejection. No implicit extrapolation.

`Oracle`
: A non-probabilistic validator (simulation, proof, rule engine) independent of the model.

Resulting architectural rule. When a proposal controls execution, the decision boundary must be binary and externally checkable. Optimizing a probabilistic evaluator against another probabilistic generator does not bound worst-case harm.

Executive summary of the two results

Lemma 1 (Residual Error Inevitability): For any finite model deployed in an environment with non-zero probability of encountering novel-support inputs, there exist inputs where the model's error is unbounded by internal confidence or calibration. A worst-case bound on error cannot be achieved by scaling or confidence alone.

Theorem 1 (Utility–Harm Separation): Expected-value decision criteria (including expected utility, VaR/CVaR variants) cannot bound worst-case harm under novelty. A worst-case harm bound collapses into an explicit admissibility constraint: either an action is admissible or it is not.

---

1. Residual Error Inevitability Lemma

Setting. Let X denote the space of possible inputs (states, contexts, specifications). A deployed system receives samples from a distribution P over X. Let f be a target mapping (ground truth) and h be the model's predictor. Define a loss $l(h(x), f(x)) \in [0, 1]$.

Lemma 1. Suppose deployment includes novelty: there exists a set $N \subset X$ such that $P(N) > 0$ and N is not covered by the effective training/validation support in a way that would uniquely identify f on N. Then, for any deterministic predictor h, the worst-case loss on N satisfies $\sup_{x \in N} l(h(x), f(x)) = 1$ for some admissible f. For randomized rules, the worst-case expected loss is at least 1/2.

Proof sketch. Consider two possible target functions f_0 and f_1 that agree on all observed/trained regions but disagree on a novel point x^* in N. Any fixed predictor h must output a single value at x^*. That output matches at most one of the two targets. Therefore there exists an admissible world (f_0 or f_1) in which the loss at x^* is maximal. This is an information-theoretic ambiguity, not a capacity limitation. Randomization cannot remove the ambiguity; it can only average between the two possibilities.

Interpretation for execution systems. Novelty implies there is no universal guarantee that high confidence correlates with correctness in the worst case. A safety argument that relies on model confidence, calibration, or even self-consistency remains probabilistic, and therefore cannot provide a hard pre-execution boundary.

Corollary. If a system must guarantee that certain failures never occur (or occur with a worst-case bound), then the guarantee cannot be delegated to the model's internal uncertainty estimates. It must be enforced by an external admissibility condition that is checkable without referencing the model's beliefs.

---

2. Utility–Harm Separation Theorem

Problem. Many safety approaches treat risk as a penalty term in an objective: maximize expected utility minus expected harm. This works when the distribution is known and stable. Under novelty, it does not bound worst-case outcomes.

Definitions. Let action a selected by the system induce an outcome with harm $H(a, x) \geq 0$ for state x. A decision rule uses a risk functional ρ over random variables (e.g., expectation, VaR, CVaR, spectral risk). We ask: does minimizing $\rho(H)$ bound ess sup H (the worst-case harm over events of non-zero

probability)?

Theorem 1. For any risk functional ρ that depends on average or tail statistics (including expectation, VaR, CVaR, and mixtures), there exist distributions with novelty support such that a rule optimizing ρ permits actions with arbitrarily large worst-case harm. The only functional that directly bounds worst-case harm is the essential supremum, which is equivalent to enforcing an explicit admissibility constraint $H(a, x) \leq H_{\max}$ for all admissible x.

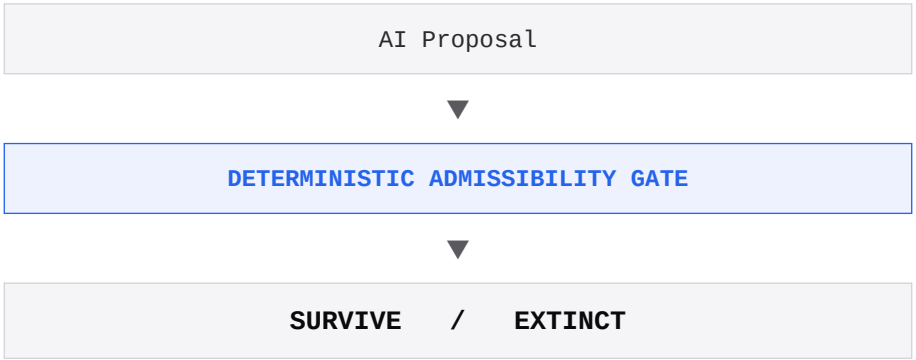
Proof sketch. Construct a family of distributions where a catastrophic event has small but non-zero probability ε. Expected-value and tail-average criteria can be made insensitive by letting ε shrink while letting catastrophe magnitude grow. Thus the optimized objective can remain low while worst-case harm diverges. Bounding the essential supremum requires stating and enforcing a hard constraint that prohibits the catastrophic region entirely.

Interpretation. When execution is irreversible or safety-critical, "optimize risk" is the wrong abstraction. The correct abstraction is: define admissibility (constraints), then allow only admissible actions. In such systems, utility is not a scalar score to trade off against harm; harm boundaries are structural.

---

3. FCAL as a Structural Consequence

The two results above motivate an architectural separation:



Principle. The gate is independent of the model. It is implemented as an external oracle: physics simulation, formal proof, domain code checks, or bounded rule systems. The gate decides admissibility; the model proposes candidates.

Fail-closed rule. Under-specification and uncertainty are treated as non-admissible. The gate does not extrapolate. It either produces evidence of admissibility under stated assumptions, or it blocks execution.

Model-bound discrepancy disclosure. A critical case is when the validator is correct within its capability profile but the proposal implicitly assumes phenomena outside that profile. The correct response is neither PASS nor "LLM is wrong," but a qualified outcome with an explicit epistemic

boundary disclosure.

Minimal architecture (implementation-neutral)

Component	Role
Generator	Proposes candidate actions/designs (LLM/agent).
Classifier	Routes high-stakes / physical / regulated queries to validation.
Oracle	Evaluates against physics or formal limits under a capability profile.
Gate	Issues SURVIVE/EXTINCT (or qualified reject) before execution.
Audit log	Stores inputs, capability profile, evidence, and decision for compliance.

What this is not. This is not a policy guardrail, not a prompt template, and not an LLM evaluating an LLM. It is a deterministic execution boundary tied to non-negotiable constraints.

---

4. Domain Examples

The admissibility boundary is domain-agnostic; the oracle is domain-specific. Below are minimal examples of the same gate applied to three domains.

Structural — Eurocode admissibility

Verdict: EXTINCT	
Check	**Buckling utilization > 1.0**
Computed ratio	**1.42**
Governing mode	**Flexural buckling, y-axis**
Standard	**EN 1993-1-1 §6.3.1**

Oracle computes capacity checks per code. If required parameters are missing, the run is rejected fail-closed.

Circuit — Kirchhoff + thermal envelope

Verdict: EXTINCT	
Check	**Power dissipation exceeds rating**

Computed	**57.6 W vs 0.125 W rating**
Thermal limit	**ΔT exceeds IPC-2152 envelope**
Standard / law	**Kirchhoff + IPC-2152**

Oracle enforces electrical laws and rating envelopes. This is not "is the text plausible"; it is feasibility and safety.

Governance — risk mandate bounds

Verdict: EXTINCT	
Check	**Exposure exceeds authorized limit**
Proposed exposure	**€4.2M vs €2.5M limit**
Mandate	**Max 15% single-sector concentration**
Proposed allocation	**27% single sector**

Oracle enforces mandate bounds and approval requirements. The gate rejects actions outside authority regardless of narrative quality.

---

5. Implications

Why this layer is missing. Production AI systems have strong incentives to ship optimistic behavior quickly. Most current guardrails are probability-shaping mechanisms (filters, rankings, preference models). They reduce average error, but they do not implement a worst-case admissibility guarantee under novelty.

What changes with a deterministic gate

1) Execution becomes auditable. Each decision has evidence and a capability profile. The verdict is reproducible.

2) Liability becomes separable. The model proposes; the boundary decides. Responsibility is architecturally distinct.

3) Systems become composable. The gate stays the same; only the oracle changes. New domains require new validators, not new architectures.

Practical note on modeling limits. A validator validates a model of reality, not reality itself. Therefore, admissibility must always be conditional on an explicit capability profile and must disclose model-bound

mismatches when assumptions diverge.

References

- EN 1993-1-1: Design of steel structures — General rules and rules for buildings (Eurocode 3).
- IPC-2152: Standard for Determining Current-Carrying Capacity in Printed Board Design.
- Kirchhoff's circuit laws (KCL/KVL).
- FCAL interpretive airgap architecture and capability-profile gating (provisional filing context, USPTO).