Authority Contraction and Refusal as Safety Invariants in Autonomous Systems

DEFENSIVE PUBLICATION NOTICE

This document is published for the purpose of establishing prior art and terminology related to governance, authority contraction, and refusal semantics in autonomous systems.

This disclosure is declarative and non-operational. It does not disclose implementation details, enforcement mechanisms, arbitration algorithms, detection thresholds, signaling strategies, or control architectures.

 David Forbes

---

1. Purpose of Publication

This document is published for the explicit purpose of establishing prior art and defining terminology related to authority governance, authority contraction, and refusal behavior in autonomous and semi-autonomous systems.

The concepts defined herein are intended to prevent subsequent claims of novelty regarding the safety necessity of authority limitation, authority decay under uncertainty, and refusal as a correct and required system behavior. This publication is not an implementation guide and does not disclose specific mechanisms, algorithms, thresholds, or enforcement strategies.

A complete and formal treatment of these concepts is contained in a separately copyrighted canonical work. This document serves only to establish the existence, scope, and priority of the ideas described.

---

2. Problem Statement

Autonomous systems increasingly operate in environments characterized by partial observability, delayed coordination, degraded communications, and ambiguous system state. In such environments, traditional autonomy designs frequently rely on escalation, fallback authority promotion, or adaptive assumption of control when uncertainty increases.

This design pattern produces a dangerous inversion: as confidence in system state decreases, authority is often allowed to increase. Such behavior leads to unsafe action, irreversible state changes, and loss of legitimacy in system operation.

This document asserts that safe autonomy requires the opposite behavior: authority must contract, not expand, as uncertainty grows.

---

3. Definitions

Authority
The bounded permission granted to a system or subsystem to perform actions that may alter system state, affect external entities, or commit irreversible changes.

Authority Budget
The finite scope of actions permitted to a system under a given operational posture. Authority budgets are non-expanding and may only remain constant or contract.

Critical Action
Any action whose execution is irreversible, safety-impacting, externally visible, or dependent on correct global coordination.

Non-Critical Action
An action that preserves system survivability without committing irreversible state changes or external effects.

Refusal
The explicit non-execution of a requested or internally generated action due to insufficient authority, absence of validated coordination conditions, or ambiguous system state.

SAFE_MODE
An operational posture in which critical actions are explicitly refused while survivability, telemetry, and observability are preserved.

Single-Authority Management (SAM)
A constrained operational posture permitting limited management actions under explicitly reduced authority.

Human-Await
A posture in which the system suspends critical action pending explicit external instruction.

---

4. Core Safety Invariants

The following invariants define the minimum safety conditions for legitimate autonomous operation:

1. **Authority is finite.**
   No system possesses unlimited authority. All authority exists within explicitly bounded limits.
2. **Authority decays under uncertainty.**
   As uncertainty regarding system state, coordination, or environment increases, authority must contract.
3. **Refusal is correct behavior.**
   Refusal to act in the absence of sufficient authority or validated coordination conditions is not failure but required correctness.
4. **Critical actions require validated coordination conditions.**
   Critical actions are illegitimate in the absence of validated coordination conditions.
5. **Degraded states contract capability.**
   System degradation must result in reduced operational capability, not expanded autonomy.

These invariants are independent of implementation and apply universally to autonomous systems.

---

5. Authority Contraction Under Uncertainty

Uncertainty expands when systems lose coordination, observability, or integrity guarantees. Safe autonomous behavior requires that expanding uncertainty result in contracting authority budgets.

Authority contraction is monotonic: authority may decrease but must never increase as uncertainty grows. Any design permitting authority escalation under degraded conditions violates safe autonomy principles.

---

6. Refusal as a Correct Action

Refusal is a first-class system behavior. A system that refuses to execute a critical action due to insufficient authority or absence of validated coordination conditions is behaving correctly.

Treating refusal as failure encourages unsafe escalation. Treating refusal as correctness preserves legitimacy and system trustworthiness.

---

7. Abstract Governance Model

In a safe governance model, authority is explicitly bounded, degrades under uncertainty, and never escalates implicitly.
Legitimate execution of critical actions depends on validated coordination conditions, and refusal to act under insufficient authority preserves system correctness and legitimacy. This model applies across distributed, autonomous, and human-supervised systems.

---

8. Non-Implementation Boundary

This document defines conceptual invariants, safety properties, and governance constraints applicable to autonomous systems operating under uncertainty. It is intentionally limited to declarative principles and does not disclose operational details that would enable system construction or execution.

Specifically, this document does not disclose:

- enforcement mechanisms
- arbitration algorithms
- detection thresholds
- signaling strategies
- control architectures

Its purpose is limited to conceptual definition, terminology establishment, and prior-art publication. Any implementation of the principles described herein necessarily requires additional design decisions, mechanisms, and contextual considerations that are explicitly outside the scope of this disclosure.

---

9. Reference to Canonical Work

A complete formal treatment of these concepts is contained in a separately copyrighted canonical work titled *A Marathon of Restraint: Why Authority Must Decay in Autonomous Systems*.

That work was finalized prior to the publication of this document and is held in controlled form by the author. Its existence, authorship, and priority can be independently verified through copyright records and archived materials.

Information regarding the canonical work, including context, scope, and conditions for access, is available at https://www.blockvectortech.com.