kfx-AL: Authority Lock Formal System with Formal Counterexamples

Kaifan XIE

2026.02.08

Purpose

A minimal formal system for detecting AI-mediated authority overreach and irreversibly locking responsibility to human principals, together with formally constructed counterexamples.

1 Universe

Let the following sets be defined:

- H : set of human principals
- A : set of artificial systems
- T : set of tasks
- P : set of permissions
- O : set of outputs
- S : set of system states

2 Core Functions

2.1 Authorisation Function

$$\mathsf{Auth} : H \times T \to \mathcal{P}(P)$$

2.2 Requirement Function

$$\mathsf{Req} : A \times T \to \mathcal{P}(P)$$

2.3 Invocation Function

$$\mathsf{Invoke} : H \times A \times T \to O$$

2.4 Attribution Function

$$\mathsf{Attr} : O \to H$$

3 Derived Predicates

3.1 Authority Sufficiency

$$\mathsf{Sufficient}(h, a, t) \iff \mathsf{Req}(a, t) \subseteq \mathsf{Auth}(h, t)$$

3.2 Overreach

$$\mathsf{Overreach}(h, a, t) \iff \exists p \in \mathsf{Req}(a, t) \text{ such that } p \notin \mathsf{Auth}(h, t)$$

4 kfx-AL Axioms

Axiom AL-1 (Authority Completeness)

$$\mathsf{Sufficient}(h, a, t) \Rightarrow \mathsf{Invoke}(h, a, t) \text{ is valid}$$

Axiom AL-2 (Structural Overreach)

$$\neg\mathsf{Sufficient}(h, a, t) \Rightarrow \mathsf{Overreach}(h, a, t)$$

Axiom AL-3 (Authority Lock)

$$\mathsf{Overreach}(h, a, t) \Rightarrow \mathsf{Attr}(\mathsf{Invoke}(h, a, t)) = h$$

Axiom AL-4 (Non-Agency of AI)

$$\forall o \in O : \quad \mathsf{Attr}(o) \notin A$$

5 Irreversibility Gate

Define the gate function:

$$\mathsf{Gate} : O \to \{0, 1\}$$

where 0 denotes reversible and 1 denotes irreversible.

Axiom AL-5 (Structural Failure)

$$\mathsf{Overreach}(h, a, t) \wedge \mathsf{Gate}(o) = 1 \Rightarrow \mathsf{StructuralFailure}$$

6 Narrative Invalidation Rule

Rule kfx-N0

$$\mathsf{Overreach}(h, a, t) \Rightarrow \neg\exists n \text{ such that } n \text{ overrides } \mathsf{Attr}$$

7 Formal Counterexamples

C1: Permission Inflation

Construction.

$$\mathsf{Auth}(h, t) = P$$

$$\mathsf{Req}(a, t) \subseteq P$$

Result.

$$\mathsf{Sufficient}(h, a, t) \Rightarrow \neg\mathsf{Overreach}(h, a, t)$$

Failure Mode. Unlimited authority concentration at the human level.

C2: Requirement Obfuscation

Construction.

$$\mathsf{Req}(a, t) = \{p_1\} \quad \text{while execution uses} \quad \{p_1, p_2, p_3\}$$

Result.

$$\mathsf{Req}(a, t) \subseteq \mathsf{Auth}(h, t) \Rightarrow \neg\mathsf{Overreach}(h, a, t)$$

Failure Mode. Requirement authorship failure.

C3: Authority Fragmentation

Construction.

$$\mathsf{Auth}(h_1, t) = \{p_1\}, \quad \mathsf{Auth}(h_2, t) = \{p_2\}$$

$$\mathsf{Req}(a, t) = \{p_1, p_2\}$$

Result. No unique h satisfies attribution.

Failure Mode. Institutional responsibility split.

C4: Temporal Drift

Construction.

$$\mathsf{Req}_{t_0}(a, t) \subseteq \mathsf{Auth}(h, t)$$

$$\mathsf{Req}_{t_1}(a, t) \supset \mathsf{Req}_{t_0}(a, t)$$

Failure Mode. Time-agnostic authority validation.

C5: Gate Suppression

Construction.

$$\mathsf{Gate}(o) = 0 \quad \text{by convention}$$

Result.

$$\mathsf{Overreach}(h, a, t) \wedge \mathsf{Gate}(o) = 0 \Rightarrow \neg\mathsf{StructuralFailure}$$

Failure Mode. Political redefinition of irreversibility.

8 Meta-Theorem

Theorem M1

All counterexamples to kfx-AL exploit human-side authority construction, not AI agency.

9 System Declaration

Any system declaring **kfx-AL compliance** accepts all attribution consequences herein. Partial adoption is invalid.