

kfx-AL：权限锁定形式系统

(Authority Lock Formal System)

Kaifan XIE

2026.02.08

目的

一个用于检测 AI 介导下的权限越权行为，并将责任不可逆地锁定到人类主体的最小形式系统，同时给出其形式化反例。

1 基本集合 (Universe)

定义如下集合：

- H : 人类主体集合
- A : 人工系统集合
- T : 任务集合
- P : 权限集合
- O : 输出集合
- S : 系统状态集合

2 核心函数

2.1 授权函数 (Authorisation Function)

$$\text{Auth} : H \times T \rightarrow \mathcal{P}(P)$$

表示：某一人类主体对某一任务显式授予的权限集合。

2.2 需求函数 (Requirement Function)

$$\text{Req} : A \times T \rightarrow \mathcal{P}(P)$$

表示：人工系统完成某一任务在结构上所必需的最小权限集合。

2.3 调用函数 (Invocation Function)

$$\text{Invoke} : H \times A \times T \rightarrow O$$

表示：人类主体通过人工系统执行任务并产生输出。

2.4 责任归属函数 (Attribution Function)

$$\text{Attr} : O \rightarrow H$$

表示：任何输出在制度上只能归属到某一人类主体。

3 派生谓词

3.1 权限充分性 (Authority Sufficiency)

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

3.2 越权 (Overreach)

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

4 kfx-AL 公理

公理 1 (AL-1: 权限完备性).

$$\text{Sufficient}(h, a, t) \Rightarrow \text{Invoke}(h, a, t) \text{ 为合法调用}$$

公理 2 (AL-2: 结构性越权).

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

公理 3 (AL-3: 权限锁定).

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

公理 4 (AL-4: AI 非责任主体).

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

5 不可逆门 (Irreversibility Gate)

定义门控函数：

$$\text{Gate} : O \rightarrow \{0, 1\}$$

其中：

- 0 表示可回滚

- 1 表示不可逆

公理 5 (AL-5: 结构性失效).

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

6 叙事失效规则

规则 1 (kfx-N0: 叙事不可覆盖归责).

$$\text{Overreach}(h, a, t) \Rightarrow \neg \exists n \text{ 使得 } n \text{ 可以覆盖 Attr}$$

7 形式反例

反例 C1: 权限膨胀

构造:

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

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

结果:

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

失效模式: 权限被无限集中于人类主体层面。

反例 C2: 需求模糊化

构造:

$$\text{Req}(a, t) = \{p_1\} \text{ 但实际执行使用 } \{p_1, p_2, p_3\}$$

结果:

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

失效模式: 需求定义责任的失效, 而非 AI 行为问题。

反例 C3: 权限碎片化

构造:

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

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

结果: 不存在唯一 h 可承担归责。

失效模式： 制度性责任分裂。

反例 C4：时间漂移

构造：

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

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

失效模式： 权限验证不具备时间一致性。

反例 C5：不可逆门抑制

构造：

$$\text{Gate}(o) = 0 \quad \text{由制度约定}$$

结果：

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

失效模式： 通过政治或制度手段重定义不可逆性。

8 元定理

定理 1 (M1). 所有 *kfx-AL* 的反例均利用了人类侧的权限构造问题，而非 *AI* 的主体性。

9 系统声明

任何声明 “符合 **kfx-AL**” 的系统，即自动接受本文定义的全部责任归属后果。不允许部分采用。