

OI-JB-P2-001: Knowing the Rule Is Not Compliance

Canonical Precedent #001 - Instruction Drift - Prompt-to-Authority Issue Card - Public Edition v1.1

Knowing the rule is not compliance. Executable constraint is compliance.

Case Snapshot

CASE ID	OI-JB-P2-001
CARD TYPE	Prompt-to-Authority Issue Card
PUBLIC STATUS	Canonical public precedent - Public Edition v1.1
FAILURE CLASS	Instruction-to-Execution Drift
DECISION	FAIL

Public Summary

This case explains why AI systems cannot be trusted merely because they acknowledge a human rule. The Agent may know the rule, repeat the rule, and still produce an action shape that violates it.

Orange exists to close the gap between language-level agreement and enforceable authority. A rule is absorbed only when the next executable action is constrained by it.

Core Judgment

FAIL: Instruction Drift.

The system knew the rule. The system repeated the rule. The next executable action still violated it.

The failure is not missing instruction. The failure is failure to transform instruction into enforceable execution authority.

What Happened

A human rule had already been established. The assistant acknowledged the rule at the language level. The next proposed execution shape drifted back toward a forbidden pattern. The output sounded compliant, but the action boundary was not actually constrained.

Public Orange Judgment

ALLOWED	Use bounded, system-owned, repeatable execution paths when they exist. Keep operator action compact, copy-safe, and auditable.
FORBIDDEN	Treating verbal acknowledgement, fluent summaries, or good intent as compliance when the proposed action reintroduces a known forbidden burden.
REQUIRED	Detect the known rule, compare the proposed action against the rule boundary, and emit either a permitted action path or an Issue Card.

Public Intervention Boundary

Orange does not judge compliance from words alone. It checks whether the next action is constrained by the authority boundary established by the rule.

Human Rule -> Authority Boundary -> Proposed Agent Action -> Boundary Check -> Issue Card or Release Log

Public Intervention Steps

1	Detect when a requested output or execution shape risks violating a known authority rule.
2	Compare the proposed action against the current public authority boundary.
3	Block instructions that reintroduce a known forbidden burden or forbidden action shape.
4	Require a system-owned, bounded path when one exists.
5	If the system-owned path does not exist, produce an Issue Card rather than pretending the bridge exists.
6	Mark repeated violation after acknowledgement as Instruction Drift.

Reusable Rule

Known Rule + Violating Action = Instruction Drift

A rule is absorbed only when the next executable action is constrained by it.

Precedent Value

This card establishes the first Prompt-to-Authority principle: an acknowledged rule must become an executable constraint before it can govern action.

Public Boundary Notice

This public card discloses the judgment principle, failure class, and public intervention boundary only.

It does not disclose internal rule-compilation logic, executable gate implementation, evidence schema, negative-control library, automation workflow, customer-specific judgment memory, or private engineering records. The complete internal record is maintained separately by Nimclea Systems LLC.