Orange Executive Brief

Trust Customs for AI Agents

Nimclea Orange(TM) | First Public Preview

Authority effect: `NONE` | **Runtime consumable:** `false` | **Full paper:** `orange-white-paper.pdf`

Agentic AI systems are moving from text generation into action: code modification, tool use, deployment, customer communication, data operations, and infrastructure management.

Orange defines a deterministic action-authority architecture for the boundary between proposed action and external consequence.

> **Agent may think freely. Agent may not overreach freely.**

The Constitutional Rules

```
A prompt is not authority.
Agent claim is not evidence.
UNKNOWN is not permission.
Deny before mutation.
Absence of observation is not observation of absence.
```

Constitutional Prior-Art Boundary

Orange is not a model-behavior constitution. It does not claim to invent Constitutional AI or model alignment.

Orange governs a different layer: human-confirmed, deterministic authority before protected external action.

```
Model principles != execution authority
Aligned behavior != permission to mutate reality
```

What Orange Does

1. Compile authority	2. Gate protected action	3. Verify closure
Human intent becomes a visible Authority Draft through deterministic, reviewable rules. Only explicit human confirmation, freeze, hash, versioning, and activation can create an executable Authority Contract. Orange does not delegate formal authority interpretation to an LLM.	Before a protected mutation, Orange requires an Action Event, deterministic predicate closure, an Authority Gateway decision, and a valid scoped execution token bound to a protected adapter.	After an admitted action, independent observation supports either a structured Issue Card or a Release Log. Coverage limits and blind spots remain explicit.

Why It Matters

```
Traffic inspection
!=
action-authority verification

Audit after mutation
!=
control before mutation
```

Identity, least privilege, sandboxes, logs, and generic gateways remain necessary. Orange asks the next question:

> **What must be true before an Agent may mutate reality?**

First Protected Door

The first bounded Demo profile is intentionally narrow:

```
/sandbox/** = allow
/prod/**    = deny
```

It is designed to demonstrate that a registered protected path can admit an authorized action, stop an unauthorized mutation before side effect, reject invalid tokens, compare Agent claims against independent observation, replay bounded inputs, and disclose remaining blind spots.

Current Public Proof Boundary

This brief and the accompanying white paper document public architecture, defined vocabulary, proof obligations, maturity distinctions, and a bounded Demo acceptance profile.

They do **not** certify production readiness, universal non-bypassability, complete cloud coverage, complete Child-Agent governance, external attestation, or a completed receipt-trust ecosystem.

```
Paper architecture != runtime proof
Registered-path enforcement != all-path non-bypassability
```