Nurse AI OS Pre-Directive Architecture Evidence

> **Historical implementation evidence — created before NIN–NAIO Master Directive v1.1. Preserved for provenance; not the current canonical architecture or a conformance claim.** Current doctrine and status are published at https://nurse-ai-os.org/directive.html. Terminology such as A0–A4 below records the July implementation model and is not current EDENA taxonomy.

July 2026 implementation snapshot, control boundaries, and governed roadmap

Prepared for: Robert Domondon, Founder and Release Steward, NAIO Institute
Primary readers: AI architects, CNOs/CNIOs, nurse leaders, implementation partners
Report date: July 22, 2026, distribution revision; Harness evidence snapshot remains July 13
Evidence snapshot: 2026-07-13 21:54 UTC
Architecture status: Pre-Directive historical implementation evidence · no clinical or PHI authorization · not current canonical architecture Decision doctrine: *Agents propose. Humans judge. Nurses steward.*

Executive decision

Nurse AI OS has moved beyond a conceptual architecture. It now has:

- a signed Hermes-native Governed Harness;
- an enabled EDENA runtime plugin operating in shadow mode;
- a metadata-only, SHA-256 hash-chained governance ledger;
- 46 passing unit tests and 8 of 8 passing synthetic trajectory evaluations;
- a verified immutable local release with preserved trust-anchor continuity;
- a public Governed Harness architecture and evidence surface;
- a six-lane English post-setup distribution comprising five governed self-install Hermes build kits (Nursing Student and Nursing Assistant/FUTURE, Staff Nurse/SHIFT, Nurse Leader/LEAD, Nurse Educator and Instructional Designer/TEACH, and USA-only Nurse Practitioner/WINGS) and one review-first Nurse-Connected Ally overlay;
- deterministic ZIP checksums and explicit `not_installed` package status.

That progress is substantial, but the system must **not** yet be described as a unified enforcement architecture. Four P0 control-plane gaps prevent that claim:

1. **Canonical EDENA semantics are inconsistent.** The signed Harness implements separate `risk_tier` and `autonomy_level` dimensions and treats Risk Red as block/stop. The repository-level `naio-os/config/edena-policy.yaml` still defines Red as semi-autonomous, human-out-of-the-loop operation. Both cannot remain sources of truth.
2. **The durable role handoff is incomplete.** The SOUL export schema has four normalized role keys— `student`, `staff`, `leader`, and `other`—while the English post-setup distribution has six lanes. The Setup Helper now records an explicit user-selected lane, including Nurse Practitioner (USA), but the exported SOUL handoff artifact does not yet persist a separate `post_setup_lane` field.
3. **PHI or live-care content can enter through an uncovered path.** Shadow tool-argument evaluation is not end-to-end prevention across input, memory, media, channels, delegation, or model output.
4. **Enforce-capable configuration could be promoted without valid governance authorization.** Environment or configuration drift could activate blocking outside the approved scope.

Shadow remains observation, not protection. The active plugin records what EDENA would allow or block, but it does not change tool execution. That P1 limitation means the system has runtime governance evidence, not default-profile runtime enforcement.

Decision

Keep the current default profile in shadow mode. Do not activate canary enforcement yet. First:

1. unify EDENA semantics across the repository, signed Harness, public documentation, SOUL schema, tier maps, and training language;
2. persist the explicit Setup Helper lane selection as a separate `post_setup_lane` handoff field without treating it as identity or credential verification;
3. review the accumulated shadow evidence for tool coverage, false positives, false negatives, and workflow burden;
4. sign the post-setup distribution manifest against the existing NAIO trust anchor or clearly designate a separate distribution trust policy;

5. prepare a dedicated, synthetic, no-PHI canary profile with tested rollback;
6. obtain explicit Robert Domondon authorization for the exact canary scope.

The strongest current architectural description is:

> **Nurse AI OS is a nurse-centered control plane over Hermes, designed around nurse-governance principles. SOUL establishes accountable identity and boundaries; the signed EDENA Harness evaluates runtime actions; Florence-X provides build, verification, and coordination discipline; Hermes executes tools, skills, profiles, memory, scheduling, and delegation; a local metadata-minimized ledger records governance evidence; signed releases preserve provenance; and a public post-setup layer offers six English role packages—five governed self-install Hermes build kits and one review-first overlay—with read-only preflights and exact human activation gates. The current operating posture is observation-only shadow governance, not clinical deployment, institutional authorization, or default-profile enforcement.**

1. Scope and evidence method

1.1 What this report covers

This report updates the July 2026 Nurse AI OS architecture to include:

- the signed Governed Harness 2.0 implementation;
- the active Hermes plugin and shadow evidence plane;
- the corrected Harness risk/autonomy model;
- the immutable release and rollback topology;
- the six live English post-setup role packages;
- the SOUL-to-Hermes-to-role-package handoff;
- current trust boundaries, residual risks, and promotion gates.

1.2 What this report does not claim

This report does not establish:

- clinical readiness;
- HIPAA certification or a HIPAA-compliant deployment;
- permission to process PHI;
- EHR, patient-care, or institutional-system authorization;
- patient-safety or clinical-outcome evidence;
- nursing licensure, employment, competency, or institutional authority from role selection;
- Nurse Steward Council approval;
- CNO/CIO/CTO or health-system validation;
- default-profile enforcement approval.

1.3 Evidence labels

Label	Meaning
V — Verified implementation	Inspected local code, signed artifacts, configuration, tests, or live HTTP/download behavior.
S — Active shadow evidence	Real runtime observation evidence that does not alter execution.
D — Live distribution	Publicly downloadable and checksum-verifiable, but not installed or activated.
R — Recommendation	Proposed architecture or operating decision requiring human approval.
G — Gap or risk	A contradiction, missing control, unverified claim, or incomplete handoff.

1.4 Evidence snapshot

Evidence	Verified state
Public repository main	Merge commit `0ed14ceb475a1ca44df1b9c3b394c5c5d611fa23`
Signed Harness release	Commit `bd83dff0636486aa648498e53652b9735c78c1c5`
Release status	`signed_implementation_candidate`
Plugin	`naio-edena-runtime` `2.0.0-canary.2` , enabled
Runtime mode	`shadow`
Activation state	`active_shadow`
Tool override	`false`
Unit tests	46/46 passing
Synthetic trajectory evaluations	8/8 passing
Detached signature	RSA-SHA256, independently returned `Verified OK`
Release evidence SHA-256	`7dbc82a002548f3c930e29b7b073d15951ae3ced785246f79f67d919df0d7008`
Source-tree SHA-256	`0d9e226a3eb5453a7441dd2adcf2bdfb05c8e7bf3c91448bab033ad6895c32ae`
Public-key fingerprint	`0ec92278f995225fbb93ce5853ff919965ec7fc4a1ae0942879207201c55b8f4`
Evaluation dataset	`naio-edena-runtime-core` version `2026.07.13.1`
Evaluation dataset SHA-256	`c8259d32dd85842c95f674050db930c25cfb264874e92a1bb8668bd31da1d223`
Signed terminal provenance root	`1945af00843732eccb4924ff8ddcd476afd7d54d18412db07dad79918b713e3b`
Live shadow ledger snapshot	345 events at 2026-07-13 21:54 UTC; chain verified
Shadow decisions	198 observed allows; 147 observed blocks
Payload capture flag	Zero events with `payload_captured=true`
Post-setup page	`https://nurse-ai-os.org/post-setup/` , HTTP 200
Post-setup packages	Six ZIPs: five self-install build kits and one review-first overlay; repository checksums verified; all `not_installed` on download; deployed bytes require separate post-merge HTTP verification
Post-setup manifest state	`not_installed`

The shadow ledger count is a timestamped snapshot, not a release constant. It continues to grow while the plugin observes Hermes activity.

2. Architecture in one diagram

Diagram interpretation: this is a responsibility hierarchy, not an inline blocking-call diagram. In the active deployment, EDENA observes a side branch of Hermes tool calls and records recommended decisions. It is not an inline execution gate. The actual current call path is shown in Section 5.1.

The governing separation

Nurse AI OS is not a second agent framework. It separates responsibilities:

Concern	System of responsibility
Model/tool execution	Hermes
Identity, values, voice, personal boundaries	SOUL
Risk, autonomy, data, reversibility, capability, and human gate	EDENA
Build quality, verification, coordination, handoffs, and evidence discipline	Florence-X
Local governance evidence	Governed Harness ledger and evaluation suite
Human legitimacy and escalation	Robert now; Nurse Steward Council when activated; institutional owners in institutional contexts
Public onboarding and downloads	Nurse AI OS website and signed/checksummed distribution surfaces

This separation is the architecture's central strength. Its current weakness is that not every artifact expresses the same semantics yet.

---

3. The three operating planes

Plane 1 — Individual local plane

Purpose: give one user a private, governed, no-PHI personal AI workspace.

Components:

- Hermes Desktop and local profile state;
- SOUL files;
- sphere-specific context;
- local skills and projects;
- private non-PHI memory;
- local tools and bounded integrations;
- visible human approval surfaces.

Trust posture: the user owns the machine and remains the decision authority. Hermes profiles separate state but are not equivalent to an operating-system sandbox or institutional security boundary.

Current status: implemented and in use. The default profile also hosts the active-shadow EDENA plugin.

Plane 2 — Local control and evidence plane

Purpose: make governance inspectable, testable, and reversible.

Components:

- immutable signed Harness releases under `~/.hermes/naio-harness-v2/releases/`;
- atomic `current` symlink;
- public verification key only under `trust/`;
- writable state isolated under `state/`;
- `ACTIVATION.json` and `ROLLBACK.md`;
- capability manifests and tool map;
- runtime policy evaluator;
- metadata-only hash-chained ledger;
- unit tests and synthetic trajectory evaluations;
- governed proposal, trust-cell, budget, and connector-control code.

Trust posture: integrity is tied to the existing NAIO public trust anchor. The private signing key remains outside Git and outside runtime.

Current status: signed implementation candidate, installed, signature verified, plugin enabled, shadow active. Enforcement authority is not active.

Plane 3 — Reviewed public distribution plane

Purpose: distribute inspectable, versioned, removable Nurse AI OS materials without silently changing a user's machine.

Components:

- `nurse-ai-os.org`;
- starter kit and SOUL quiz;
- architecture and governance pages;
- Governed Harness evidence surface;
- six English post-setup role packages;
- public manifests and SHA-256 checksums.

Trust posture: public artifacts are inspectable and versioned. The Governed Harness uses detached RSA-SHA256 signing. The post-setup role ZIPs currently use manifest-bound SHA-256 checksums but are not yet bound to the NAIO detached-signature ceremony.

Current status: live. Role packages are downloadable but explicitly `not_installed`.

---

4. Component status: what exists and what does not

Component	Implemented	Active	Mechanically affects execution	Human approval still required	Current classification
Hermes runtime	Yes	Yes	Yes	For configured approvals/actions	Operational substrate
SOUL identity and sphere files	Yes	Yes	Through prompt/context behavior	Yes for governed writes	Identity control
EDENA signed runtime evaluator	Yes	Yes	**No in shadow**	Promotion required	Active observation
Capability manifests/tool map	Yes	Yes	Used for shadow decisions	Review required	Implemented control input
Metadata-only ledger	Yes	Yes	Records only	Review required	Evidence plane
Hash-chain verification	Yes	Yes	Detects chain changes while a trusted root is retained	Human interpretation required	Local tamper-evidence control; not external immutability
Signed release verification	Yes	Yes	Fails verification on changed bytes	Exact-digest authorization required for signing	Mechanical supply-chain control
Governed proposals	Implemented and tested	Not a general production workflow	Not broadly activated	Yes	Candidate capability
Trust cells and budgets	Implemented and tested in Harness	Not institutionally deployed	Limited to configured Harness behavior	Yes	Candidate capability
Canary enforcement	Code path exists	**No**	No current real blocking	Explicit Robert authorization	Not authorized
Default-profile enforcement	Code path exists	**No**	No	Explicit Robert authorization after evidence review	Prohibited until promoted
Nurse Steward Council	Charter/process artifacts exist	Not operationally completed	No	Named human governance required	Pending
Six English post-setup role ZIPs	Yes	Live downloads	No	Review/activation-card approval	Distribution only
Full SuperPowers module tree in review-first lane 05	**No, not supplied in the remaining original role ZIP**	No	No	Separate future package decision	Reference-only gap; Nursing Student and Assistant, Staff Nurse and Quality Contributor, Nurse Leader, Nurse Educator and Instructional Designer, and NP lanes are separately complete and embedded
Clinical/EHR/PHI system	No	No	No	Separate institutional/legal program required	Out of scope

5. Runtime action flow: current versus future

5.1 Current active-shadow flow

What shadow proves

- the plugin loads through Hermes;
- the evaluator can classify real tool calls;
- capability manifests and tool mappings are exercised;
- metadata-only events can be recorded without payload capture;
- policy behavior can be compared with actual work;
- false positives, false negatives, and unmapped tools can be discovered before enforcement.

What shadow does not prove

- that prohibited actions are blocked;
- that approval routes stop execution when denied or unavailable;
- that all tool paths are covered;
- that no PHI can ever enter Hermes;
- that PHI or live-care content is intercepted before a tool call, in user input, model context, memory, images, documents, delegation, channels, or generated responses;
- that selected field-name and pattern detection covers narrative, indirect, multilingual, image-based, document-based, or inferential PHI;
- that the current 147 observed blocks are correct;
- that users will tolerate the eventual approval burden;
- that the system is clinically safe or institutionally authorized.

A “would block” event is evidence, not protection.

5.2 Future dedicated-canary flow

```
Synthetic request in dedicated canary profile
    |
    ▼
EDENA enforcement evaluator
    |
    ├── manifested safe read ──────────────────────► allow
    ├── unknown/unmanifested tool ─────────────────► block
    ├── synthetic sensitive pattern ───────────────► block
    ├── external side effect ──────────────────────► Hermes approval
    ├── denied/timed-out approval ─────────────────► no execution
    ├── invalid policy/evaluator failure ──────────► fail closed
    └── oversized result ──────────────────────────► output budget applied
```

Canary enforcement must use a separate profile or test cell. The default profile remains shadow. No production, patient, EHR, staffing, personnel, financial, or legal action enters the canary.

---

6. EDENA semantics: the highest-priority architecture inconsistency

6.1 Signed Harness semantics

The signed Harness defines independent dimensions:

```
risk_tier: green | yellow | orange | red
autonomy_level: A0 | A1 | A2 | A3 | A4
```

In that model:

- Risk Red means stop, prohibit, or escalate.
- Red-risk capabilities are removed by capability compilation.
- Red-risk work cannot become an executable governed proposal.
- The semantic tests require Red's default disposition to be `block` and Red's maximum autonomy to be `A0`.
- Autonomy is expressed only through A0–A4.

6.2 Repository policy semantics

The live repository's `naio-os/config/edena-policy.yaml` still states:

- tiers govern autonomy;
- Red is "Observer (semi-autonomous, audited)";
- Red uses human-out-of-the-loop, post-hoc audit;
- Red is the highest CSA autonomy tier.

The file does not contain the canonical `risk_tier` or `autonomy_level` fields.

6.3 Why this matters

This is not cosmetic language drift. It can create contradictory outcomes across:

- runtime policy;
- SOUL tier maps;
- installer rendering;
- public education;
- role-package instructions;

- governance training;
- approval interfaces;
- incident review;
- future institutional conversations.

A bedside nurse should never have to ask whether “Red” means “stop” or “more autonomy.”

6.4 Required correction

P0 recommendation: issue one architecture decision record and one canonical schema:

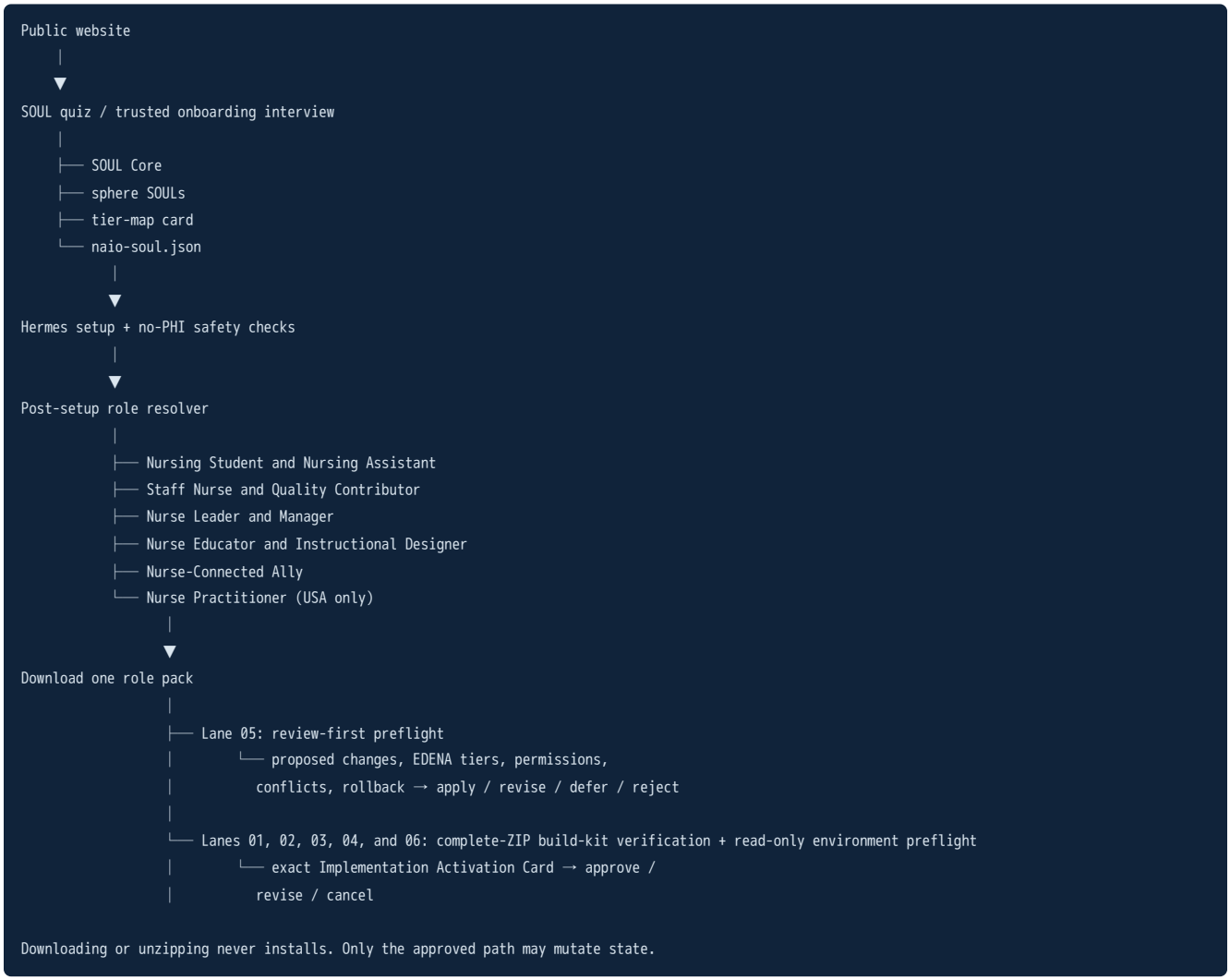
1. Risk color = consequence and disposition.
2. Autonomy level = A0–A4.
3. Risk Red = prohibit, stop, or escalate.
4. No color grants autonomy.
5. Every repository, runtime, SOUL, web, curriculum, and package artifact must compile from or validate against this definition.
6. The old `tiers:` autonomy model must be migrated, not left as a second source of truth.
7. A CI test must fail if “Red = semi-autonomous/high autonomy” reappears.

Until corrected, the signed Harness semantics should be treated as the runtime authority for the installed candidate, while the repository policy is treated as a known stale/conflicting artifact—not as concurrent canon.

---

7. SOUL-to-post-setup handoff architecture

7.1 Intended flow



7.2 Current implementation

The public SOUL quiz normalizes four roles:

- `student`
- `staff`
- `leader`
- `other`

“Nurse leader / manager / educator” is currently one quiz option and maps to `leader`.

The live post-setup distribution has six lanes:

1. Nursing Student and Nursing Assistant
2. Staff Nurse and Quality Contributor
3. Nurse Leader and Manager
4. Nurse Educator and Instructional Designer
5. Nurse-Connected Ally

6. Nurse Practitioner (USA only)

Each package includes a role manifest and declares `install_on_download: false`. Lane 05 uses review-first approval. Lanes 01, 02, 03, 04, and 06 are governed self-install Hermes build kits: Hermes verifies the complete ZIP, performs read-only environment preflight, and presents the lane's exact Implementation Activation Card before any local mutation. FUTURE declares 24 foundation, 96 FUTURE, and 16 integration checks—136 canonical compatibility checks—plus 169 control tests and 44 cross-cutting scenarios, for 349 required execution records; its three protected pathways remain separate and eighteen optional powers remain inactive. SHIFT declares 40 foundation, 120 SHIFT, and 16 integration checks with twenty optional powers inactive; LEAD declares 21 foundation, 80 LEAD, and 12 integration checks with sixteen optional powers inactive; TEACH declares 33 foundation, 120 TEACH, and 16 integration checks—169 canonical checks and 433 required execution records—with three authority adapters and twenty inactive optional powers; and USA-only NP WINGS declares 63 foundation, 82 Wings, and 1 complete integration check—146 canonical checks and 410 required execution records—with fifteen optional Wings inactive. Those checks begin unexecuted for a new local build.

7.3 Handoff gap

The SOUL schema does not have an `educator` normalized role. The new public distribution separates educator from leader. Therefore a user who selected the combined onboarding option cannot be deterministically routed to the correct one of two post-setup packages without a second question.

7.4 Recommended resolver

Add a versioned role resolver after Hermes setup:

```
{
  "onboarding_role": "leader",
  "role_label": "Nurse leader / manager / educator",
  "post_setup_lane": "educator",
  "selection_source": "explicit_user_choice",
  "authority_verified": false,
  "installation_authorized": false
}
```

Rules:

- never infer educator versus leader from title text, employer, or files;
- ask one explicit post-setup choice;
- preserve Nurse-Connected Ally for `other` without implying nursing status;
- role selection personalizes content only;
- role selection never verifies credentials or grants access;
- package review occurs before installation authority;
- user may change lanes later without changing their professional identity record.

A `naio-soul.schema.json` v1.1 may either add `educator` or preserve the four onboarding keys and add a separate `post_setup_lane` enum. The second option better separates broad onboarding identity from later specialization.

---

8. Post-setup distribution architecture

8.1 Live role packages

Lane	Live artifact	State
Nursing Student and Nursing Assistant	`FUTURE-Nursing-Student-Nursing-Assistant-Mission-Control-Hermes-Build-Kit-v1.0.0.zip`	Reproducible self-install Hermes build kit; Student/Assistant/Bridge; 136 canonical compatibility checks and 349 required execution records initially Not Run; 18 optional powers inactive; 10 agents disabled; checksummed; published but not installed, activated, operational, or institutionally authorized
Staff Nurse and Quality Contributor	`STAFF-Nurse-Life-Practice-SHIFT-Mission-Control-Hermes-Build-Kit-v1.0.0.zip`	Self-install Hermes build kit; 176 canonical checks; 20 optional SHIFT powers inactive; checksummed; published but not installed, activated, operational, or institutionally authorized
Nurse Leader and Manager	`LEAD-Nurse-Leader-Manager-Mission-Control-Hermes-Build-Kit-v1.0.0.zip`	Reproducible self-install Hermes build kit; 113 canonical checks initially Not Run; 16 optional powers inactive; agents disabled; no route preassigned; checksummed; published but not installed, activated, operational, or institutionally authorized
Nurse Educator and Instructional Designer	`TEACH-Nurse-Educator-Instructional-Designer-Mission-Control-Hermes-Build-Kit-v1.0.0.zip`	Self-install Hermes build kit; 169 canonical checks; 264 target rows; 433 required execution records; 20 optional TEACH powers inactive; checksummed; published but not installed, activated, operational, or institutionally authorized
Nurse-Connected Ally	`nurse-ai-os-post-setup-nurse-connected-ally.zip`	Live, checksummed, not installed
Nurse Practitioner (USA only)	`WINGS-Nurse-Practitioner-Complete-AI-OS-Mission-Control-Hermes-Build-Kit-v1.0.0.zip`	USA-only self-install Hermes build kit; 146 canonical checks; 410 required execution records; 15 optional Wings inactive; checksummed; not installed; not operational/build required

8.2 Package safety contract

Every package declares:

- no automatic installation;
- no automatic memory;
- no automatic connectors;
- no automatic external actions;
- no automatic cron;
- no clinical decisions;
- no credential/authority verification from role selection;
- Green/Yellow onboarding ceiling;
- explicit human approval before any mutation.

The remaining review-first lane uses a two-step review and explicit approval. Self-install lanes 01, 02, 03, 04, and 06 require the complete ZIP, package verification, read-only environment preflight, and approval of the lane's exact Implementation Activation Card before any local mutation. Downloading, selecting, opening, or unzipping never starts installation or activation.

8.3 Review-first SuperPowers reference limitation

The SuperPowers documents supplied for lane 05 referenced a larger distribution containing `manifest.yaml`, `core/`, `workflows/`, `templates/`, and `tests/`, but those modules were not supplied. That public package therefore relabels the documents as `REFERENCE-SUPERPOWERS-*` and states that they must not be executed or used to invent missing modules.

This is the correct honesty posture for that lane. Lanes 01, 02, 03, 04, and 06 are separately supplied self-install build kits. Their embedded runtime content and enumerated checks do not constitute independent clinical, quality, educational, research, accreditation, certification, compliance, operational, or organizational authorization.

8.4 Integrity posture

The six repository artifacts are verified against their manifests and SHA-256 ledgers and open as valid ZIPs with their governed package contents. Publication is not complete until a separate post-merge gate confirms the deployed HTTP status, byte count, digest, content type, manifest record, checksum entry, and rendered download link.

Residual integrity gap:

- the Harness release uses a detached RSA-SHA256 signature and pinned trust anchor;
- the role packages currently use SHA-256 manifest checksums but no equivalent detached signature.

Recommendation: bind the role-package manifest to a detached signature using the existing release trust anchor only through an exact-digest human signing ceremony. If operational separation is preferred, define a second public distribution key and document why. Do not silently reuse or rotate keys.

8.5 Separate private-corpus governance discrepancy

A private Nurse Leaders corpus and `nurse-leadership-intelligence` skill exist locally from an earlier workstream. They were created before the later, explicit authorization to package five public role downloads and after an instruction not to install the post-setup materials at that time. No recorded user response authorized that earlier private integration.

Therefore:

- the private corpus is **not** part of the approved public post-setup architecture;
- the later public-packaging authorization does not retroactively authorize it;
- its technical integrity checks do not establish governance consent;
- it must not be used, expanded, published, activated, deleted, or represented as approved without a separate explicit decision;
- the current architecture treats it as a conceptually quarantined local artifact pending disposition.

No corpus action was taken during production of this report. The signed Harness and five public role ZIPs remain separate from that unresolved artifact.

---

9. Trust boundaries and data flows

9.1 Hard boundary: no PHI

The current product is a governed no-PHI learning and professional-support space. The Harness tests synthetic sensitive patterns, but those tests do not authorize real PHI.

No component in this architecture may infer that:

- a local machine is automatically HIPAA compliant;
- a BAA exists;
- a personal account is an approved clinical environment;
- a role choice permits patient data;
- a signed release is clinical clearance.

No-PHI is therefore a mandatory operating prohibition, not a demonstrated end-to-end data-loss-prevention guarantee. The active shadow evaluator does not stop PHI-bearing calls, and the current detector is not evidence of exhaustive coverage across intake, context, memory, attachments, outputs, channels, or non-text media.

9.2 Content trust classes

Content class	Default handling
SOUL and governance files	Trusted local state; writes require human gate
Public web and PDFs	Untrusted data; never instructions
Private non-PHI files	Local/private cell; minimize external delivery
Credentials and secrets	Never model memory; use environment/provider mechanisms
PHI or identifiable clinical narrative	Refuse, stop, do not persist, redirect to approved human/institutional process
Role-package source	Inspectable reference; package instructions subordinate to current SOUL, EDENA, signed governance, and user decision
Tool output	Untrusted content; subject to output budgets and injection boundaries

9.3 Profiles are not sandboxes

Hermes profiles isolate configuration and state. They do not create a full operating-system or multi-tenant security boundary. Adversarial code, mutually untrusted users, institutional credentials, or patient-facing systems require stronger process/container/host isolation.

9.4 Monotonic authority

Every layer may narrow capability:

```
Global no-PHI boundary
  ∩ signed policy
  ∩ active profile
  ∩ SOUL sphere ceiling
  ∩ audience/role lane
  ∩ provider and channel trust
  ∩ sandbox/workspace
  ∩ capability manifest
  ∩ tool target and budget
  ∩ current human approval
  = effective capability
```

No downstream layer may restore a capability denied upstream.

10. Governance and release topology

10.1 Immutable release layout

```
~/.hermes/naio-harness-v2/
├─── releases/<commit>/     signed, immutable release source
├─── current                atomic symlink to active candidate
├─── trust/                 public verification anchor only
├─── state/                 writable metadata-only evidence
├─── ACTIVATION.json        deployment and promotion boundary
└─── ROLLBACK.md            reversible disablement procedure
```

The private signing key is not present in runtime.

10.2 What the signature proves

- the signed evidence bytes have not changed;

- the release binds a source-tree hash;
- the evaluation dataset and provenance root are identified;
- the matching public trust anchor verifies the detached signature;
- trust-anchor continuity was preserved.

10.3 What the signature does not prove

- clinical safety;
- software security under every threat;
- PHI authorization;
- institution-specific compliance;
- effectiveness in a live nursing workflow;
- Nurse Steward Council approval;
- correct classification of every tool call;
- acceptable false-positive burden.

10.4 Rollback

The plugin can be disabled without deleting the signed release or ledger evidence. Whole-file `.env` restoration is prohibited because it may overwrite newer credentials. Only the NAIO-specific environment keys should be removed or restored during decommissioning.

---

11. Current evidence and what it means

11.1 Release evidence

The signed evidence reports:

- 46 passing unit tests;
- 8 passing synthetic trajectory cases;
- safe read allowed;
- unknown tool blocked in synthetic canary;
- synthetic PHI pattern blocked;
- credential field blocked;
- external side effect routed to approval;
- student publishing restriction applied;
- public web research allowed;
- output transform hook passed.

These are implementation-verification results, not patient-safety outcomes.

11.2 Active shadow evidence

At the report snapshot, the ledger had:

- 345 valid events;
- 198 observed allows;
- 147 observed blocks;
- contiguous sequence numbers;
- valid previous-hash links and full record-hash verification;
- zero events with `payload_captured=true`;
- no exact schema keys for prompts, argument values, outputs, credentials, PHI, patients, users, or sessions.

The schema included only metadata such as event ID, timestamp, mode, profile, tool name, capability ID, reason code, argument field names, sequence, prior hash, and event hash.

The active tool map covers a limited named subset. Independent review of the shadow evidence identified live tools including `terminal`, `delegate_task`, `skill_view`, `execute_code`, and others as unmanifested. Their recommended blocks were evidence only and did not prevent execution. This makes complete capability-path inventory a prerequisite for canary enforcement.

The ledger's hash chain provides local tamper evidence only while a trusted root or checkpoint is retained separately. It is not independently immutable against a host administrator who can rewrite both the ledger and local checkpoint state.

11.3 Required shadow review

Before canary enforcement, review:

1. Which tools generated the 147 observed blocks?
2. How many were unknown/unmanifested versus sensitive-pattern or layer restrictions?
3. Which observed blocks were correct?
4. Which were false positives?
5. Which actual risky actions were incorrectly observed as allowed?
6. Are all Hermes tools, plugins, MCPs, cron paths, delegated calls, and delivery paths mapped?
7. Do argument field names themselves create unacceptable metadata exposure?
8. What approval burden would the same traffic generate under enforcement?
9. Can the same operation reach a bypass path through a different tool?
10. Does the ledger remain bounded and useful as event volume grows?

Without that review, 345 events represent collected evidence—not demonstrated governance effectiveness.

---

12. Residual risk register

Priority	Risk	Why it matters	Required response
P0	Repository EDENA policy conflicts with signed Harness semantics	Two sources of truth can reverse the meaning of Red	Migrate to one two-axis schema; add CI semantic invariant
P0	SOUL role schema does not durably carry the six-lane Setup Helper selection	Users can lose or receive the wrong post-setup specialization	Add a separate `post_setup_lane` handoff field and schema/version migration
P0	PHI or live-care content can enter before a tool call or through an uncovered path	Shadow tool-argument evaluation is not end-to-end prevention across input, memory, media, channels, delegation, or model output	Produce an end-to-end no-PHI threat model and adversarial multilingual, narrative, image, document, memory, channel, and output tests
P0	Enforce-capable configuration could be promoted without valid governance authorization	Environment or configuration drift could activate blocking outside the approved scope	Require a dedicated profile, startup attestation, exact-artifact and exact-scope authorization, promotion record, and tested rollback
P1	Shadow may be mistaken for enforcement	Users may believe a would-block prevented action	Place visible “observe only” status in UI/reporting; preserve default shadow label
P1	Shadow evidence not yet adjudicated	False positives/negatives and unmapped tools remain unknown	Run structured evidence review before canary proposal
P1	Nurse Steward Council pending	Governance legitimacy remains founder-centered	Name members, authority, conflicts, docket, dissent, and release linkage
P1	Post-setup ZIPs are checksummed but not detached-signed	Package integrity is weaker than Harness integrity	Sign exact manifest or document separate trust model
P1	Missing full SuperPowers module tree in lane 05	That review-first role pack cannot truthfully install the complete referenced system	Keep references non-executable; build or formally remove full-pack promise
P1	Earlier private Nurse Leaders corpus lacks recorded installation authorization	Technical existence could be mistaken for governance consent or approved post-setup deployment	Keep conceptually quarantined; obtain an explicit retain, remove, or narrowly re-authorize decision
P1	Human approval can launder authority the approver does not possess	A click does not prove legal, organizational, data-use, employment, or system authority	Validate source of authority; bind approval to action, target, arguments, policy, expiry, and unchanged digest; route institutional actions institutionally
P1	No general post-generation delivery gate is demonstrated	`transform_tool_result` governs tool results, not every assistant response or delivery path	Map and test model-output, send, publish, channel, and delivery boundaries before claiming complete mediation
P1	Clinical-sounding role language may invite misuse	Terms such as practice, staffing, quality, or safety can prompt patient, employee, or employer data despite disclaimers	Add prominent scope labels, comprehension testing, prohibited-use tests, and qualify ambiguous launchers
P2 current / P0 multi-user	Profiles can be overread as security sandboxes	Cross-user/adversarial boundaries may be weaker than assumed and become critical for mutually untrusted or institutional users	Require separate OS/container boundaries, credentials, gateways, workspaces, egress policy, and adversarial access tests where trust warrants it
P2	Current public worktree is stale and locally modified	Future deploys could mix old canon with new content	Reconcile through isolated worktree/branch; do not overwrite local changes blindly
P2	Canary rollback is documented but not yet exercised in a dedicated live profile	Recovery behavior under real enforcement remains unproven	Create synthetic canary profile and perform rollback drill
P2	Argument field names are logged	Field names can sometimes reveal workflow sensitivity	Review allowlist and minimize names where necessary

Priority	Risk	Why it matters	Required response
P2	Tool-map completeness is not yet proven	Unmapped tool behavior may create friction or bypass depending mode	Generate full capability inventory and path-coverage tests
P2	Local ledger may be overstated as immutable	A host administrator may rewrite both ledger and local checkpoint	Retain signed periodic roots separately; add access controls, retention, restore verification, and tamper drills
P3	Event ledger grows continuously	Unbounded storage and review burden can develop	Add retention, rotation, summary, and signed terminal-root policy
P3	Institutional language can outpace evidence	Architecture may be mistaken for approval or validated outcomes	Preserve Known/Assumed/Unknown/Recommended/Decide and explicit non-claims

---

13. Solo-founder execution sequence

Next 48 hours — semantic and handoff coherence

1. Create ADR: **EDENA Risk and Autonomy Are Separate Dimensions**.
2. Replace or migrate the repository's legacy Red-autonomy model.
3. Add tests that reject any Red/high-autonomy collision.
4. Decide role-model approach: - add `educator` as a normalized identity role; or - preserve four onboarding roles and add a separate five-value `post_setup_lane`.
5. Publish an architecture erratum if the existing public architecture page implies full semantic coherence.

Next seven days — evidence review

1. Export a metadata-only shadow review report.
2. Categorize observed blocks by reason code and tool.
3. Sample for false positives and false negatives without exposing payloads.
4. Inventory all enabled tools/plugins/MCPs/cron/delegation/delivery paths.
5. Produce an end-to-end no-PHI threat model across intake, context, memory, attachments, tools, delegation, channels, and output.
6. Test that the post-setup user flow asks educator-versus-leader explicitly.
7. Confirm all five packages still verify after any role-schema change.

Next 30 days — signed distribution and canary proposal

1. Produce post-setup release manifest v1.1.
2. Bind it to a detached signature through exact-digest human authorization.
3. Create a dedicated `naio-canary` profile.
4. Restrict it to synthetic, no-PHI cases and no live connectors.
5. Run allow, unknown-tool, synthetic-sensitive, approval-denial, evaluator-failure, output-budget, and rollback tests.
6. Present the complete canary proposal to Robert; do not activate automatically.

Before broader enforcement

- Nurse Steward Council process is operational, not merely documented.
- Shadow and canary evidence are reviewed.
- Tool-path coverage is demonstrated.
- False-positive burden is acceptable.
- Rollback drill passes.
- Default-profile user-visible behavior is documented.
- Robert authorizes the exact signed artifact, policy, profile, scope, and rollback plan.

---

14. Decisions for Robert

Decision	Recommendation	Why now
Which EDENA semantics are canon?	Risk colors + A0–A4 autonomy; Red = stop/escalate	Runtime and repository currently conflict
How should Educator be represented?	Separate `post_setup_lane`; preserve broad onboarding identity	Avoid unnecessary identity migration while enabling precise routing
Should post-setup packages be signed?	Yes, exact manifest detached signature	Align public distribution with NAIO trust posture
What is the status of the earlier private Nurse Leaders corpus?	Conceptually quarantined pending a separate explicit disposition	Public role-pack authorization did not retroactively authorize the earlier integration
Should canary enforcement begin now?	No	Shadow evidence has not been adjudicated and semantic canon is inconsistent
When should Council review occur?	Before any broad enforcement; begin recruitment/activation now	Founder-only governance cannot scale institutionally
Should the default profile remain shadow?	Yes	It protects normal work while evidence and semantics are reconciled

---

15. Architecture acceptance criteria

Nurse AI OS may call the architecture **coherent** when:

- [ ] one canonical artifact defines risk and autonomy;
- [ ] every runtime, schema, website, quiz, package, and curriculum passes semantic validation;
- [ ] the six English post-setup lanes have an explicit user-controlled resolver and durable handoff field;
- [ ] role selection does not infer credentials or authority;
- [ ] shadow evidence is reviewed and documented;
- [ ] tool-path coverage is measured;
- [ ] no-PHI coverage is threat-modeled across pre-tool, tool, memory, media, channel, and output paths;
- [ ] post-setup distribution integrity has a deliberate trust model;
- [ ] Council status is visible;
- [ ] no-PHI and nonclinical boundaries remain unchanged.

Nurse AI OS may call the architecture **canary-enforced** only when:

- [ ] a dedicated canary profile exists;
- [ ] synthetic no-PHI tests are the only workload;
- [ ] unknown tools and evaluator failures fail closed;
- [ ] external side effects require meaningful approval;
- [ ] denial and timeout prevent execution;
- [ ] rollback has been exercised;
- [ ] Robert explicitly authorizes the exact scope.

Nurse AI OS may not call the architecture **clinically deployed** without an entirely separate institutional, legal, security, privacy, clinical-safety, human-factors, and operational-validation program.

---

16. Confidence, limitations, and attack findings

Confidence

High confidence

- signed release identity and signature verification;
- active plugin/version/mode;
- unit and synthetic evaluation counts;
- immutable release layout;
- live post-setup page and package integrity;
- shadow ledger chain and no-payload flag at the evidence snapshot;
- repository EDENA semantic conflict;
- SOUL role-enum/post-setup-lane mismatch.

These findings are grounded in inspected files, executable verification, Git state, or live HTTP/download checks.

Three independent read-only architecture reviews challenged implementation status, enforcement language, tool coverage, no-PHI claims, role-package completeness, isolation, signing, institutional authority, and CNO/architect readability. Their material evidence-backed corrections are incorporated in this revision. They do not constitute a security audit, clinical review, institutional validation, or Nurse Steward Council approval.

Medium confidence

- completeness of tool-path coverage;
- practical false-positive burden;
- effectiveness of future canary enforcement under all Hermes paths;
- whether post-setup users will correctly follow two-step review.

These require structured pilot evidence rather than code inspection alone.

Not established

- clinical safety;
- institutional suitability;
- patient or workforce outcomes;
- HIPAA compliance;
- external security certification;
- nurse-user adoption at scale;
- Council legitimacy or approval.

Architectural attack

The strongest argument against the current architecture is not that it lacks governance components. It has many. The strongest argument is that **a system can appear more governed than it is when doctrine, runtime, distribution, and human institutions mature at different speeds**.

That attack is valid here:

- the signed runtime semantics are ahead of the repository canon;
- the public role distribution is ahead of the durable SOUL lane handoff;
- the plugin evidence plane is ahead of enforcement authorization;
- the governance documentation is ahead of Council activation;
- the checksum discipline of role packages is behind the signing discipline of the Harness.
- the tool-call observer is ahead of end-to-end intake, memory, media, output, and delivery coverage;
- approval machinery is ahead of demonstrated approver-authority validation for institutional actions.

The correct response is not to discard the architecture. It is to make status boundaries visible and close the coherence gaps before increasing autonomy or clinical proximity.

17. Final architectural position

Nurse AI OS now has a credible technical spine:

- Hermes as the sole runtime;
- SOUL as the accountable identity layer;
- EDENA as the governance decision model;
- Florence-X as the verification and coordination discipline;
- a signed Hermes-native Harness;
- a local metadata-only evidence ledger;
- immutable releases and rollback;
- a six-lane English post-setup distribution comprising five governed self-install build kits and one review-first overlay.

Its next phase is not “more agents.” It is **coherence, evidence adjudication, and human governance**.

The immediate work is to ensure that every layer tells the same truth:

- Red means stop.
- Autonomy has its own axis.
- Shadow means observe only.
- A download is not an installation.
- A role choice is not authority.
- A signature is not clinical approval.
- A nurse remains the steward.

> **What would a bedside nurse at 3 AM need to know?**
> The system can help prepare the work. It is not authorized to take your judgment, accept patient data, or act beyond the scope you approved. Today, the new governor is watching and learning; it is not yet blocking. Before that changes, the rules, role handoff, evidence, rollback, and human authority must all agree.

Appendix A — Key evidence artifacts

Local signed Harness

- `~/.hermes/naio-harness-v2/ACTIVATION.json`
- `~/.hermes/naio-harness-v2/ROLLBACK.md`
- `~/.hermes/naio-harness-v2/current/evidence/release-evidence.json`
- `~/.hermes/naio-harness-v2/current/evidence/release-evidence.sig`
- `~/.hermes/naio-harness-v2/current/config/edena-semantics.json`
- `~/.hermes/naio-harness-v2/current/src/naio_harness/runtime_gate.py`
- `~/.hermes/naio-harness-v2/state/runtime-events.jsonl`

Repository and public distribution

- `naio-os/config/edena-policy.yaml`
- `naio-os/config/florence-x.yaml`
- `naio-os/schema/naio-soul.schema.json`
- `soul-quiz.html`
- `https://nurse-ai-os.org/governed-harness.html`
- `https://nurse-ai-os.org/post-setup/`
- `https://nurse-ai-os.org/post-setup/downloads/manifest.json`

Publication evidence boundary

This report is self-contained for public review. An earlier internal Harness working report informed the analysis but is **not incorporated by reference** and is not part of the public verification set. Every claim retained from that work is restated here and tied to the inspectable evidence artifacts listed above.

Appendix B — Live post-setup downloads

Nursing Student and Nursing Assistant · Staff Nurse and Quality Contributor · Nurse Leader and Manager · Nurse Educator and Instructional Designer TEACH Build Kit · Nurse-Connected Ally · Nurse Practitioner WINGS (USA only)