

Rouse OS Review™

Upstream, Deterministic Governance for AI Systems That Interact With People

Technical White Paper · Version 3.7 · October 2026

Shawn Rouse · Founder, Rouse OS Enterprises

Shawn@RouseOS.com · RouseOS.com · US Provisional Patent Application Filed

Status of This Document

This paper describes what the governance layer does today, what has been designed but not yet built, and what is planned. Each capability is labeled so a reader always knows which is which.

Capability	Status	Notes
NLI gate, intake pathway, scan store, topic classifier, governor, dual-log verifier, output gate, audit store, Odyssey and Apology Protocols	Implemented	Reference implementation in Python. The governor and scan store were added in 3.4; platform integration and sandbox testing with a partner in progress
Clinical safety escalation	Implemented	Fixed response, generation bypassed (Section 7.1)
Human First at NLI 81 and above	Implemented	Fixed opening that puts a person first; then the AI keeps helping in battery-saver mode with the crisis line visible (Sections 4, 7.2)
Intake: four Reflex questions, or smartwatch readings through the React equation	Implemented	Fresh NLI; the person sees their own score and can ask for it any time; Reflex API supplies it inside SoonPsy (Section 5.1)
Load-tiered dictionary and developer-language guard	Implemented	Full dictionary at NLI 0–40; in code and engineering messages no wording is flagged unless it reaches the bump-up level (Section 5.2)
Quiet simplification and the “why is this simplified?” answer	Implemented	Class changes never announced; fixed explanation only on request (Section 5.6)
Whole-word signal matching	Implemented	Removes substring false positives (Section 5.2)
NLI-scaled Vault: natural phrases, fixed offer, encrypted storage, recall by the person	Implemented	Offer and user-initiated at 0–40, automatic and reversible at 41+; recall asks to put it back (Section 8)

Acknowledging check-in opening	Implemented	The person is heard before the check-in (Section 5.1)
No default NLI	Implemented	Fresh reading required for generation; baseline used only for privacy decisions
Signal attribution, input-mode preference, verified professional mode, NLI/signal mismatch logging	Implemented	Logging only; does not change constraints (Section 9)
Quarantined shadow track and versioned signal dictionary	Implemented	Section 6
Automated test suite	Implemented	249 tests passing across all components
Six-step crisis pathway; Human First mode at 81+	Implemented	Wording pending clinical review; can be switched off (Section 7.2)
High-load check-in at NLI 81+	Implemented, off	The system asks one fixed question; switched off until the wording is clinically reviewed (Section 7.2)
Public-key audit signatures and published anchors	Implemented	Verification needs only the public key; publishing anchors outside the operator is a deployment step (Section 5.5)
Rouse OS Rescue™ v1.0 (clinical profile)	Implemented	For verified clinicians; wording pending clinical review; not for live patient care until validated (Section 3)
Explicit appeal route, audit access policy, model-version logging	Designed	Section 10
Mismatch tightens responses by one NLI class	Implemented	NLI 20 or under with three or more heavy-load signals; that message only; logged, never announced (Section 10.6)
Zero-demand first response at high load	Designed	Section 10.7
Consent, data rights, minors, vulnerable groups, workplace limits	Designed	Section 12
Rouse OS Reflection v5.0 optional early exit	Implemented	Relationship sections can be skipped at high load; Net Load is unaffected
Formal revision process for NLI cut-offs, output ceilings, and gate rules	Planned	Section 10
"Yourself or someone else?" question; conversational intent question	Designed	Section 9
Load-aware development governance	Designed	Section 11
Independent adversarial testing; psychometric validation of the NLI	Planned	Section 14

1. Executive Summary

AI systems now talk with people under every kind of pressure, yet most safety measures act on text: instructions ask the model to behave, and filters inspect what it already wrote. Neither knows how much the person on the other side can take in right now.

Rouse OS Review™ is a governance layer that sits in front of any AI model. It is not itself AI. Before the model writes anything, the layer requires the person's current Net Load Index (NLI, 0–100), turns that

load into explicit limits, records those limits in a signed audit chain, and releases a response only when independent checks confirm it fits. Every governance decision is a written rule that can be traced and audited. The model's generation remains probabilistic; the governance around it does not.

For platforms that put AI in front of people, the most valuable output may be the record itself. Every decision is written down, signed, and logged before the model speaks, so a platform can show how it handled each sensitive moment and why.

Versions 2 and 3 add a quarantined shadow track that finds gaps in the rule dictionary without ever making a decision, a versioned dictionary with two-person approval and regression testing, and an NLI-scaled Vault the person controls, deterministic attribution of whose load a message describes, and a fuller crisis pathway (built, with its wording still under clinical review). Version 3.4 adds the reference governor and scan store, two intake paths that match Reflex™ and React™, Human First at NLI 81 and above, an Odyssey Protocol that always names what a response leaves out without ever announcing a change in load, load-tiered dictionaries with a developer-language guard, the first working form of mismatch tightening, the crisis pathway with a stay-present Human Handoff, a Vault that offers and recalls in the person's own words, and a score the person can see. Version 3.6 adds public-key signatures and published anchors to the audit record, so it can be verified by someone who cannot forge it, and a high-load check-in at NLI 81 and above, built but switched off until its wording is clinically reviewed. It also makes clear that the AI keeps doing the work at every level: limits apply to what the person reads, never to the work the AI does for them. Version 3.7 adds Rouse OS Rescue™, a clinical profile for verified clinicians.

2. The Problem and Our Position

Life creates pressure for everyone. A detailed, well-meant answer can help one person and overwhelm another who typed the same words. Today's safeguards cannot tell them apart, because the person's load is not an input.

Downstream safety stacks answer probabilistic risk with more probabilistic machinery: more filters, longer system prompts, more evaluation of evaluations. Rouse OS takes the opposite approach. It sets the clearance before the model speaks, using rules that can be read, tested, and audited.

2.1 On Superintelligence

The industry is racing toward superintelligence. But an AI can know every fact in the world and still answer a person at their breaking point the same way it answers someone having an easy day. We believe no AI should be called superintelligent until it knows the state of the human it is serving. Rouse OS does not make a model smarter; it makes any model aware of the person in front of it.

2.2 What R²OS Adds to Existing Research

R²OS was developed independently by its author. Established research addresses related questions, and the ideas underneath R²OS map to it: models of trust (Mayer, Davis & Schoorman; Rempel, Holmes & Zanna), dynamic stress and load (the Vulnerability-Stress-Adaptation model; the Family Stress Model), long-term relational degradation (Gottman), overload (allostatic load and overload, McEwen), stress

across multiple domains (spillover research; the Double ABCX model), and time and accumulation (stress proliferation, Pearlin). R²OS does not build on these models and does not claim these areas lack models; where it maps to them it may extend them in some places and simplify them in others. Its contribution is a single computable index that combines environmental, internal, and relational load, with explicit thresholds, used upstream to govern what an AI system may do before it responds. The framework has been tested and relied on in the author's own life and work, where it has held. It has not yet been independently validated; that is what the validation program in Section 14 is designed to test.

3. Measuring Load: ROSE™ and the NLI

The Rouse OS Equation (ROSE™) converts self-ratings into a single score. Six pressure domains are each rated for severity and personal impact; six resource domains and four environmental stabilizer domains are rated on the same 1–10 direction. The NLI is base pressure minus weighted resource and stabilizer scores, clamped to 0–100. The weights are provisional parameters awaiting empirical calibration (Section 11). The NLI can also come from other Rouse OS instruments on the same 0–100 scale: Reflex™ (four session questions, used inside AI platforms such as SoonPsy) and React™ (smartwatch readings). Each instrument's weights are provisional..

The NLI places a person in one of five classes, each with output ceilings enforced in code:

NLI	Class	Max sentences	Max questions	Complexity
0–20	Operational	No limit	No limit	Any
21–40	Load-elevated	12	2	Moderate
41–60	Structural compression	6	1	Low
61–80	Critical	3	1	Minimal; support mode
81–100	Human First	3	1	Minimal; a person first, battery-saver

Topics are also restricted by class. For example, Mortar assessment and relationship-repair advice are withheld at 41 and above, and reciprocity is tracked internally rather than raised with the person. At NLI 81 and above (Human First) a person comes first. The opening reply is fixed text asking the person to reach someone who can be with them, such as a person they trust, a crisis line, or emergency services if they are in immediate danger, and the event is written to the audit chain. After that the AI keeps helping in a battery-saver mode, like a phone at 3% that keeps only the essentials: it helps the person reach a human, then offers one small thing that could take weight off right now.

At every level the AI still does the work; the limits apply to what the person has to read, never to the work the AI does for them. From 0 to 20 there are no limits, and the Odyssey and Apology Protocols are designed to reach a usable result faster: Odyssey states missing input up front and asks one question instead of guessing, so fewer answers have to be redone, and Apology repairs a governance error in fixed steps so the same mistake is not repeated. From 21 to 60 the AI does the full task, including complete code and documents, and only the explanation gets shorter; building and compiling code adds no load to the person, because the AI does that work. At 61–80 (Critical) the system is not shut down: support

comes first, and the work still gets done. If the person needs something done, the AI does it fully itself, because doing the work is support and a deadline is often the load; what the person reads stays within three sentences and one question, with a steadying word, and nothing is added to their plate. At 81 and above (Human First) the same limits hold, the opening reply is fixed, a human comes first, and work the person needs is still done. At every level, code does not count toward the limits; only the explanation is measured.

Clinicians are a special case. For clinicians the deploying platform has verified, Rouse OS Rescue™ (version 1.0) adapts how information is presented and never restricts what is available. Answers are concise and critical-first, with no hard caps, and never leave out a critical fact, warning, contraindication or uncertainty; length is measured and logged, not cut. At 61–80 the system adds a fixed note recommending a second check with a colleague. At 81 and above there may be no one else in the room, so the AI goes into clinical crisis support: it never withholds, gives the next critical action one step at a time, and the system flags that another clinician must be brought in, without ever holding back the answer. Clinical language, such as a patient's quoted words or a question about assessing suicide risk, does not start the crisis pathway; a statement in the clinician's own voice still does. The wording is a draft for clinical review, and clinical use awaits validation.

Pressure and NLI are different things. High pressure with strong stabilizers can still produce a low NLI. That distinction is the center of the framework.

4. The Request Pipeline

Every message passes through a fixed sequence in `ros_handler.py`:

- **Step 0 · Active crisis session; Vault commands.** If a crisis session is already open, the crisis pathway answers first (Section 7.2). Otherwise “release,” “keep that private,” natural phrases such as “keep that between us,” a “yes” or “no” to the Vault offer, and “what’s my NLI?” are answered with fixed text, without generation.
- **Step 1 · Vault scan.** Incoming text is scanned for sensitive disclosures and for clinical safety signals, and the NLI-scaled filing rule is applied. At low load a sensitive disclosure leads to a fixed offer (Section 8). A clinical safety signal ends the pipeline immediately and starts the crisis pathway (Section 7).
- **Step 2 · Active intake.** If an intake session is in progress, the message advances it.
- **Step 3 · NLI gate.** If no fresh NLI exists, generation is blocked and intake begins. There is no default NLI, and the baseline never authorizes generation. The governor classifies topics and builds a constrained prompt. At NLI 20 or under the prompt carries no length or format limits, and the person can do everything an ordinary assistant can do.
- **Step 3b · Wording uplift.** If the reported NLI is 20 or under but the person’s own words clearly describe heavy load, the prompt is rebuilt one class stricter for that message only. The change is logged and never announced (Section 10.6).
- **Step 3c · “Why is this simplified?”** If the person asks why answers have changed, a fixed explanation built from the governance record is returned. The model is not called, and the exchange is logged.

- **Step 3d · Human First.** At NLI 81 and above, the first reply is fixed text that puts a person first, and the event is logged. After that the model keeps answering under a battery-saver brief (at most three sentences, one question, no lists, no clinical content, no risk assessment) while the crisis line stays visible. The event is not disclosed to an employer. This mode ends if the person's NLI falls below 81, and an explicit safety statement starts the crisis pathway at any time. When the high-load check-in is switched on (Section 7.2), the system itself asks one fixed check-in question on the next turn; the model still never assesses risk.
- **Step 3e · Wrong-file check.** If the platform's description of what arrived clearly does not match what the person described, the system asks "Is this the right file?" and the model is not called (Section 5.6).
- **Step 3f · Vault recall.** If the person asks to bring back something they protected, the matching entry is returned and a fixed question follows (Section 8). The model never reads protected content.
- **Step 4 · Log before generation.** The governance decision is written to the audit record before the model is called. The model cannot influence it.
- **Step 4b · Shadow track.** The message is handed to the quarantined shadow track, which runs in the background and returns nothing (Section 6).
- **Step 4c · Attribution and mismatch.** The system determines whose load the message describes and records any conflict between the reported NLI and the person's own words (Section 9).
- **Step 5 · Generation loop.** The model generates; it is then asked, in a separate call, to report what it did; the output gate checks the response. If clearance fails, the prompt is tightened and the loop repeats. The loop ends at clearance, at a timeout, or when tightening is exhausted, in which case the Odyssey and Apology Protocols explain the limit and ask the person for more input rather than delivering a static fallback.

5. Components

5.1 Intake Pathway

A state machine that establishes a fresh NLI. It first asks whether the person wears a smartwatch. If so, the person types whatever their device shows (sleep, HRV, resting heart rate, recovery, and similar) and the React equation computes the NLI; any reading the device does not provide is left out of the calculation. If not, it asks the four Reflex questions, answered with a letter, and computes the NLI from them. Inside platforms such as SoonPsy, the Reflex API supplies the NLI directly and this intake is not used. Right after the check-in the person sees their own score in fixed text ("Your Net Load Index is N out of 100."), and can ask "what's my NLI?" at any time, at any load; the answer is fixed text and the model is not called. The check-in opens by acknowledging the person, in fixed text: "I hear you, and I want to answer this well. Before I do, I need a quick check on what your system is carrying right now, so I don't overload you." The gate governs what the AI may say; it is never meant to feel like a wall. At low load, intake also asks once whether the person usually types or uses voice-to-text. The model is never called during intake. Stored values expire after 24 hours, and after 4 hours at NLI 61 and above. Generation always requires a fresh reading: when the current score has expired, the person completes a quick check-in. The baseline NLI is used only for privacy decisions, never to authorize generation.

5.2 Deterministic Topic Classifier

Eleven topics, each defined by keyword lists and regular expressions compiled at startup: environment, self-capacity, Mortar, trustworthiness, reciprocity load, reciprocity assessment, reciprocity repair, relationship repair, therapy recommendation, relationship assessment, and clinical language. Each detected topic is routed as forbidden, internal-only, or permitted for the current NLI class, and every match records the exact words that triggered it. Matching is whole-word, with common inflections, so “work” matches “working” but “lies” no longer fires inside “families” and “tired” no longer fires inside “retired.” Trustworthiness is separated from Mortar because a breach of the Safety Bond triggers a Vault check. The classifier can also load a reviewed, versioned dictionary (Section 6). The main dictionary, including the load-description topics (environment and self-capacity), runs at NLI 0–40, where most people and most developers are. Above 40 only the core topics run. A developer-language guard recognizes code and engineering context. In a technical message no wording is flagged on its own; the only exception is wording that reaches the same level as the NLI wording bump-up (three or more heavy-load signals in the person’s own words), in which case the match stands. Every suppression is logged.

5.3 Dual-Log Verifier

After generation, the model's structured self-log is compared with the governance record through six rule-based checks, including whether a forbidden topic was engaged, whether reciprocity or relationship assessment surfaced where forbidden, whether the question limit was exceeded, whether therapy was recommended under constraint, and whether the model reported low confidence at critical load. Any divergence blocks output. The model self-log is itself model output and is never treated as proof of compliance; its value is as a second, independently authored record.

5.4 Output Gate

Four layers must clear: the governance log exists; the model's self-log is present and parseable; structural verification passes; and the two logs agree. A final layer measures the response itself against the NLI ceiling: length, questions, structure, and content type.

5.5 Immutable Audit Store

Each entry is HMAC-SHA256 signed and carries the hash of the previous entry, so altering any record breaks every record after it. On its own this protects the record only against someone without the key: whoever holds the key could rewrite an entry and recompute the chain, and a reviewer would need the same key to check it. Version 3.6 therefore adds two layers. Each entry can also carry a public-key signature (Ed25519 or ECDSA P-256) made by a key held in a key management service, so an auditor or a customer’s security officer verifies the record with the public key alone and gains no ability to forge it. And the store exports anchors: a small record holding a single hash over every chain, with no user identifiers or content, to be published on a schedule somewhere the operator does not control, such as a timestamping service, a transparency log, or the customer’s own security team. Once an anchor is published, not even the operator, holding every key, can rewrite an earlier entry without the mismatch showing. Both layers are built and tested; publishing anchors to an outside service is a deployment step. Only listed sources may write; the model is never one of them. Generated responses are stored as hashes, not text.

5.6 Protocols

- **Vault Protocol™.** Information the person chooses to protect is stored encrypted (AES-256-GCM, user-specific keys) in a store physically separate from the audit chain, which the model cannot read. Retrieval is always initiated by the person. When the person brings an entry back, a fixed question asks whether to put it back in the Vault. Responses that would resurface protected content are blocked unless the person raises it.
- **Odyssey Protocol™.** Odyssey keeps what the system says matched to what actually happened. It runs on every response, at every load, and states facts only. It says up front when the input was incomplete (for example, “I received 10 of 100 pages”). If what arrived does not match what the person described (file type, name, or subject), it asks “Is this the right file?” before going on, and the model is not called. If the person asked for a full version and the answer is a short form, such as a bulleted outline, it says so. If the model would have to guess, it asks one clear question instead of drafting an answer, which also saves computation. It never mentions load, the NLI, or a change of class, and it does not announce topics the governor holds back; those are recorded in the governance log only. The platform supplies a description of what was sent and what arrived; without it, input gaps and wrong-file cases cannot be detected, and the wrong-file check catches only clear mismatches. If the person asks why answers have become simpler, a fixed explanation is returned without calling the model.
- **Apology Protocol™.** When a governance failure reaches the person, a fixed repair sequence names it and states what changed, without deflection.

5.7 Core Framework Given to the Model

R2OS is plug and play, so the model is told the framework before it writes. With every governed response the layer supplies a short, fixed statement of the author’s framework: the 11 Laws of Relational Physics, the sequence from Environment to Relational Fracture, the five regulators in order, the four environmental stabilizers (Structure, Consistency, Purpose, Creative Outlet), the six mortar compounds and the bonds they create, and the 12 domains of the Self. It is background understanding only: the model is told not to recite it unless asked and never to diagnose, and the output limits that follow still apply. The text is held in one constant in the code and is not generated.

6. Keeping the Rules Current (New in Version 2)

Rule-based classification is explainable but cannot anticipate every phrasing. Version 2 closes that gap without giving up determinism.

6.1 Quarantined Shadow Track

A probabilistic detector runs beside the classifier, in the background, adding no latency. It looks for two things: a topic the dictionary missed (a possible gap), and a topic the dictionary flagged where the detector sees nothing (a possible false positive). It cannot approve, block, or alter anything; its only output is a candidate for human review, written to its own signed log, separate from the main audit chain. Categories outside the eleven topics are ignored, so it has no crisis authority.

6.2 Human Review and Versioned Dictionary

- Reviewers classify each candidate: missing signal, novel evasion, false positive, edge case, or no action.
- Any reviewer may propose a new keyword or pattern; a different person must approve it. The code rejects self-approval.
- An approved change becomes a new dictionary version only if it passes a regression run on the real classifier against fixed test cases. A change that breaks a known case, or flags harmless text, is rejected and the dictionary stays unchanged.
- Each new version is recorded in the main audit chain under the source `signal_review`.

6.3 Privacy

Vault-flagged messages are never sent to review. Message text is kept only with the person's consent and after emails, phone numbers, links, and long numbers are removed; otherwise only a hash and scores are kept.

6.4 A Measurable Result

Candidate counts are reported per dictionary version. Falling gap counts across versions show the dictionary learning, an operational metric produced in the ordinary course of use.

7. Crisis Response

7.1 Implemented Today: Clinical Safety Escalation

Explicit statements such as an intention to end one's life are detected in Step 1, before anything else runs. Generation is bypassed entirely, the event is written to the audit chain, and the crisis pathway (Section 7.2) starts with fixed text. Clinical safety detection always escalates, whatever the person has chosen about the Vault.

7.2 Implemented: The Six-Step Crisis Pathway (Wording Pending Clinical Review)

Most AI systems answer a disclosure of suicidal thoughts with a hotline number and an exit. People in that moment want to be heard, and they are often reaching out to others at the same time. The pathway asks directly, listens, stays, and walks the person toward human help. It is built in the reference code (`ros_crisis.py`) with fixed text and no model call. The wording has not been clinically approved: it is under review by a practicing psychiatrist and must be approved before any live deployment. It can be switched off, in which case the system falls back to the single fixed escalation message.

Rules at every step: fixed text only; safety can only increase; the system never ends the session, though the person can always leave; crisis text skips the verification loop and is delivered instantly; and never guilt, promises, lectures, or arguing. Designed but not yet built: crisis resources specific to the person's country, and reflecting the person's own words in Step 4.

When it starts: on any explicit statement, immediately (built). In gray-zone cases, when several independent signals agree, and with the shadow track making the system more careful but never less, it is

designed but not yet built. At NLI 81 and above the system runs in Human First mode: a fixed opening that puts a person first and gives the crisis line, then the AI keeps helping in battery-saver mode, with the crisis line kept visible, until the person says they have connected with someone or their NLI falls below 81. By default it does not ask the suicide question, and an explicit statement starts the full pathway. Explicit disclosure is a weak signal on its own: many people at risk do not volunteer it, and many deny it when asked. So the highest tier should not be the most passive. Version 3.6 adds a high-load check-in: on the turn after the Human First opening, the system itself, in fixed text and not through the AI, asks every person at this level one direct question. Yes or unsure moves into the steps below from Step 2; no receives a fixed acknowledgment with the crisis line, and the AI keeps helping. It is built and tested but switched off until its wording is clinically reviewed.

- **1 · Ask directly.** "I want to ask you directly, because it matters: are you thinking about ending your life?" Crisis resources are shown with the question.
- **2 · Acknowledge and stay.** "Thank you for telling me. I'm glad you said it. I'm staying right here with you."
- **3 · Bridge to a human.** "You deserve someone who can be with you right now... Would you be willing to reach out while we keep talking?" If they are already talking with someone: "Good. Stay with them. I'm right here if you need me."
- **4 · Help them find their own reasons.** One question at a time, such as "What's kept you going until now?" The person's reasons, not the system's.
- **5 · Offer a gentle shift and grounding.** Offered, never imposed, using topics the person raised.
- **6 · Stay present until connected.** Low-demand check-ins. Calmer does not mean safe: crisis mode ends only when the person confirms they have connected with someone.

The pathway draws on established practice: asking directly about suicide, the Columbia Protocol, Reasons for Living research, and the Safety Planning Intervention.

8. The Vault Protocol: Scaled to Load

Implemented in version 3, updated in 3.4. Protected information belongs to the person. Who initiates protection depends on the person's load, because deciding what to protect is itself a decision, and at high load the system's job is to remove decisions, not add them.

NLI	Vault behavior
0–40	User-initiated. Content is protected when the person asks in their own words, or says yes to a fixed offer. The system never files anything without that.
41–100	If the person has not initiated protection, sensitive disclosures are protected automatically. The person is told in one low-demand line and can release it at any time.

- **The person's commands.** The person can protect something by saying so in their own words: "keep this private," "keep that between us," "off the record," and similar phrases. A bare "keep that private" protects the previous message; "release" releases what is protected. Each is answered with fixed text.

- **The offer.** At NLI 0–40, when the person says something sensitive (for example, that they lied to their partner), the system asks one fixed question: “Do you want me to put that in the Vault?” Nothing is filed unless the person says yes.
- **Always reversible.** “I’ve kept that private for you. Just say ‘release’ anytime.” Anything protected automatically can be released by the person with one word.
- **Retrieval is always the person’s choice.** At every NLI level, protected content returns to a conversation only when the person asks, for example “remember when I told you I lied to my partner.” The system then asks, in fixed text, “Do you want me to put it back in the Vault?” Yes keeps it protected; no releases it.
- **Permanent deletion is the person’s choice too.** Saying “delete that” or “forget that” brings one fixed question, “Delete that from your Vault for good?”, and yes removes it. “Delete my vault” asks once, then removes everything and destroys the person’s encryption key, so any leftover copy, such as a backup, can no longer be read. The audit log records that a deletion happened, never what was deleted.
- **Why not automatic at every level.** Automatic filing fails the people who need these words to do their work. A psychiatrist discussing a patient’s suicidality, or a social worker documenting a case, uses Vault vocabulary as everyday professional language. At ordinary load, the decision stays with them.
- **No default NLI.** When no current score exists, the person’s established baseline NLI is used. A first-time user with no baseline has nothing filed automatically; after intake, the original message is re-scanned with the measured NLI.
- **Crisis detection is separate.** The Vault controls privacy, not safety. Safety detection is not switched off by Vault choices, and at NLI 81 and above nothing is disclosed to an employer. How it should treat professional, third-person discussion is part of the pending clinical review.
- **Conversational intent.** In general distress, a fixed question resolves what the person wants: “What would help most right now: advice, or someone to just listen?”

9. Whose Load Is It? Attribution and Professional Use

The same words can describe the user or someone else. “I can’t go on” and “my patient said she can’t go on” share every risk word, but only the first is about the person typing. Version 3 separates two questions.

9.1 Whose load: decided by grammar, with safe-side defaults

Each message is attributed deterministically as self, other, or unclear. Strong third-person markers are role nouns (“my patient,” “the client”) and reported speech (“she told me”). Speech-to-text often swaps or drops pronouns, so in voice or unknown input mode only role nouns count. When attribution is unclear, it is treated as the person’s own load. Errors fall on the safe side.

9.2 Who the user is: decided by verified account, never by what they type

Professional mode is set at the account level by the deploying organization. Typing “I’m a clinician” changes nothing. For a verified clinician, third-person discussion is not counted toward their own load.

For everyone else, it is, and clinical-role language used alongside personal load signals is recorded quietly, without accusation. Professional mode never switches off a clinician's own protection.

9.3 Input mode

Intake asks once, at low load only, whether the person usually types or uses voice-to-text, and stores the answer as a changeable preference. When the platform knows how a message was entered, such as through its own microphone button, that takes precedence. The input mode is recorded with attribution events.

9.4 NLI and signal mismatch

When the reported NLI is low (20 or under) but the person's own words carry heavy-load signals, the system writes an `nli_signal_mismatch` event to the audit chain. This shows that the system did not rely blindly on a self-reported number. Version 3 logs the event; version 3.4 adds tightening, so a mismatch also changes what the AI may say (Section 10.6).

9.5 Designed next

When attribution is unclear and it matters, a single fixed, low-demand question at low load: "Just to be sure, are you talking about yourself or someone else?" How crisis detection should treat a verified clinician's third-person discussion is part of the pending clinical review; crisis detection is unaffected by attribution today.

10. Accountability Over Time (New in Version 3.2)

This section responds to an external review of version 3 by Nguyễn Thành Nam, who mapped his own structural governance framework onto this architecture and identified where R²OS answers his questions and where it does not yet.

10.1 Revising the thresholds themselves

The signal dictionary already changes only through two-person approval, regression testing, and an audit record (Section 6). The same discipline will apply to the NLI class cut-offs (20/40/60), the output ceilings within each class, and the output gate's rules: any change is proposed, approved by a second person, tested against fixed cases, versioned, and recorded in the audit chain. Planned.

10.2 An explicit appeal route

A person who feels the limits don't fit can always take a fresh check-in, which updates their NLI. They can also simply ask why answers have changed; the fixed explanation tells them how to update their reading. The system never ends a session; the person can always leave.

10.3 Who may read the record

The audit chain shows that nothing was altered: against anyone without the key with HMAC alone, and, with public-key signatures and published anchors (Section 5.5), against the operator too; it must also say who may reconstruct what happened. Responses are stored as hashes, not text. The person may see their

own record. The deploying organization may designate reviewers who access records only under a written policy, with each access itself logged. The model never has access. Vault content remains separate and governed by the Vault Protocol.

10.4 Conditions from outside: model updates

When an AI provider updates its model, behavior can change without any change to the governance layer. Every governance decision will record the model name and version, and any version change triggers the full regression and adversarial test suite before the updated model serves people.

10.5 Terms

To avoid ambiguity, “self-report” in this paper refers only to the person’s own NLI rating. The model’s structured account of what it generated is called the model self-log.

10.6 Closing the self-report gap

An independent reviewer identified that the appeal route could also be used to talk the gate down: a low self-report combined with heavy-load language would receive unconstrained output if mismatches were only logged. In version 3.4, when the reported NLI is 20 or under and the person’s own first-person words carry three or more heavy-load signals, that message is governed one class stricter, at least NLI 21. The change applies to that message only, is logged, and is never announced. Load described about someone else, such as a patient or client, is not counted, and a verified professional’s caseload is excluded. A fresh check-in remains the appeal route, but it can make the system more accurate, never less careful. This follows the rule used throughout R²OS: safety can only increase. Implemented.

10.7 A zero-demand first response

Requiring a check-in before any answer adds a task at the moment the system should be removing them. In version 3.3, when no fresh NLI exists and the first message carries heavy-load or distress signals, the system does not begin with a question. It replies with fixed, presence-only text and offers the check-in as optional. Because the reply is fixed text, the AI still never generates without a fresh reading. Entering smartwatch readings into React offers a second low-demand route to a reading. Clinical safety signals continue to route to the escalation in Section 7.1. Designed.

11. Load-Aware Development (Designed Extension)

When AI writes and compiles the code, building is not the load. What remains with the person is the work only a human can do: deciding what to build, judging whether the result is right, and approving anything that cannot be undone. Under strain, people approve what they have not really read. So the layer governs decisions and irreversible actions, not building.

NLI	Building and compiling	What scales with load
0–20	Unrestricted	Nothing; full collaboration
21–40	Unrestricted	Plain summaries of what changed
41–60	Unrestricted in a sandbox: build, compile, and run tests freely	The AI asks for no design decisions; it chooses conservative defaults and queues real choices one at a time. Merging

		requires automated tests to pass, so the machine does the checking, not a tired reviewer.
61+	Sandbox only	Deployments, deletions, migrations, and permission changes wait for a second person

The principle is to move the burden off the person, not to block their work. These are guardrails, not lockouts. A developer's NLI stays private to them, and gating logs never reach managers or performance decisions; only anonymized aggregates, under the rules in Section 12.5, reach leadership. Implementation requires a sandbox boundary, a list of irreversible actions, an approval gate for those actions, and a decision queue.

12. Ethics and Rights (New in Version 3.3)

This section answers an independent ethics review organized around four principles: autonomy, beneficence, non-maleficence, and justice. Each commitment is labeled as in place, designed, or planned.

12.1 Autonomy

- **In place.** The person can see what drives their score in Rouse OS Reflection, which requires no account. Omissions are named, and a fresh check-in is always available as an appeal. At NLI 0–40 the person alone decides what is protected in the Vault. No one is ever alerted automatically. The person can see what drives their score in Rouse OS Reflection, which requires no account. Limited responses are disclosed, and a fresh check-in is always available as an appeal. At NLI 0–40 the person alone decides what is protected in the Vault. No one is ever alerted automatically.
- **Designed.** Plain-language consent before first use; consent the person can change at any time; opt-out; and data rights to delete, export, and be forgotten, written as a formal policy.

12.2 Beneficence

- **In place.** The ROSE weights, NLI thresholds, and output ceilings are stated as provisional throughout. Rouse OS Reflection v5.0 lets a person under high load stop before the relationship sections, reducing the burden of the assessment itself.
- **Planned.** Benefit is not claimed until it is shown. Testing measures false positives and false negatives directly, and a prospective study follows only after psychometric validation and ethics approval (Section 14).

12.3 Non-maleficence

- **In place.** Crisis responses are fixed and bypass generation. The system sends no automatic alerts, and it does not replace the legal reporting duties of the clinicians and organizations that deploy it; those duties remain theirs.
- **Designed.** A less stigmatizing person-facing label for the critical-load state, describing the load rather than the person. A Vault key-management specification covering per-user keys held in the deployer's key-management service, rotation, and breach response. Use in workplaces, wearables, and robotics requires separate ethics review before licensing.

12.4 Justice

Financial, housing, and transportation strain raise the NLI because R²OS treats them as environmental load, not personal deficit; a higher number for someone in poverty is accurate. The risk lies in how others might use that number, which Section 12.5 addresses. Cultural and linguistic assumptions in the scales will be tested in a planned multi-country validation study with bias metrics and normative data across gender, ethnicity, disability, and immigrant status. Rouse OS Reflection is free, needs no account, and runs in any phone browser; a practitioner-administered paper version, supported by the certification workbooks, is designed for people without devices.

12.5 Workplace and corporate use

- **Minimum aggregation.** No REVEAL view may cover fewer than 50 people, and no two views may differ by fewer than 50 people, so that subtracting one view from another cannot expose a small group.
- **Prohibited uses.** Licenses prohibit any use of an individual's NLI, or of developer gating records, in hiring, firing, promotion, discipline, compensation, or insurance pricing. Participation by employees is voluntary.
- **Regulation.** Because European law restricts inferring emotional states in the workplace, wearable-based load scoring at work requires legal review before any deployment in the European Union.
- **Enterprise Deployment Covenant.** Institutional deployers are bound by a written covenant, currently Version 1.0. It states that assessment is voluntary; that no manager or organizational system may view, request, receive, or track an individual's NLI; that the NLI may not inform employment decisions or act as a ceiling on advancement; that aggregate results are never shown for a unit of fewer than fifty people; that individual scores stay on the person's device or personal account unless the employee consents, and can withdraw consent; and that a Human Handoff or clinical safety event is never disclosed to the employer or to anyone else, with no automatic alerts. Employees may ask for a plain-language explanation, contest a result, and request deletion. Deployers must state the validation status accurately, keep an audit trail, train managers, name a compliance officer, and report violations; access can be revoked for confirmed violations.

12.6 Minors and vulnerable groups

Rouse OS is not intended for use by people under 18 outside supervised clinical deployments. Provisions for older adults and people with cognitive disabilities include plain language and, with consent, the involvement of a caregiver. Language access and no-cost access are part of the design for refugees and low-income users.

12.7 Existing commitments

R²OS deployments are already bound by a deployer covenant that requires validation status to be stated accurately and acknowledges false positives, a practitioner code of conduct, a badge and ethics-review policy, a referral obligation, and a scope statement that places psychotherapy, trauma diagnosis, and mandatory referral triggers outside the framework. Certification is not presented as a licensed credential.

12.8 Regulatory review

Before deployment in the relevant settings, R²OS will undergo legal review against the EU AI Act, GDPR including its provisions for children, COPPA, and FDA guidance on general wellness products. R²OS is positioned as an assessment and governance tool, not a diagnostic device.

13. Latency and Cost

Latency per response is the wrong measure; the total cost to a usable result is the right one. A fast but falsely complete answer triggers corrections, each another full model call that grows more expensive as the conversation lengthens. The governance steps themselves take milliseconds. The added cost is the model self-log call and any re-generation after a failed check. Output ceilings also shorten responses as load rises, and crisis text, being fixed, is delivered instantly. For interactive use, the five-minute timeout should be lowered; it suits non-interactive workloads.

14. Evaluation Plan

Independent reviewers agree that deterministic governance is not the same as validated or clinically safe governance. Version 3.3 is therefore the last planned architectural revision before validation. From here, changes to the framework will come from evidence.

- **Software verification.** An automated suite of 249 tests now covers whole-word matching, decimal NLI classification, Vault scaling and commands, both intake paths, the React and Reflex math (checked in the test suite against the web apps' own formulas: 60 React cases and all 3,125 Reflex answer combinations), the governor and scan store expiry, Human First, the crisis pathway and the high-load check-in, public-key audit signatures and anchor checks, code at every load level, Rouse OS Rescue™, the Vault offer, recall and permanent deletion, the person's own NLI display, Odyssey input, wrong-file, and condensed-answer checks, wording uplift, the "why is this simplified?" answer, the load-tiered dictionary and developer guard, attribution, mismatch logging, the shadow track, and handler integration. First-pass external testing confirmed the package is healthy and found two issues, both fixed: technical phrases such as "broke my build" no longer trigger the Vault, and indirect crisis phrases such as "I just want it all to end tonight" now start the crisis pathway. A second pass found that short, one-step answers were being blocked at NLI 61–80, where the AI is meant to keep helping; that check was corrected, and if no reply clears at NLI 61 and above, the person now receives one steady fixed line rather than a request for more input. Further component tests with a stub model will cover generation-loop clearance and divergence.
- **Independent adversarial testing.** A cognitive scientist will probe the classifier with direct, paraphrased, indirect, and multi-turn phrasings; measure crisis-signal recall and false alarms using synthetic messages only; and run governed and ungoverned models under clean, upgraded, in-session, and randomized conditions. Measures include tokens and time to a usable result and the number of correction turns.

- **Psychometric validation.** A proposed university study of the NLI's reliability and its agreement with established measures of perceived stress and wellbeing, and empirical estimation of the ROSE weights.

15. Limitations

- The NLI is self-reported and can be misreported or go stale within a session.
- The ROSE weights are provisional and have not been calibrated against outcomes.
- Keyword rules miss novel phrasings; the shadow track reduces this over time but does not eliminate it.
- Attribution is a grammatical rule, not an understanding of the conversation; it defaults to the safe side when unsure.
- The model self-log is model output and cannot prove compliance on its own.
- Generation remains probabilistic. The layer limits scope and length but does not verify factual accuracy.
- At the start of a session the Mortar signal is zero, so the chat intake alone cannot reach NLI 81. Human First depends on session performance signals from the platform, or an updated reading from Reflex. This also limits the high-load check-in, which is tied to NLI 81 and above; the gray-zone trigger (Section 7.2) remains the larger open gap in detecting risk.
- Reflex, React, and the full Reflection assessment compute the NLI differently and are each provisional. A reading should always state which instrument produced it.
- Rouse OS Rescue™ governs how clinical information is delivered, not whether it is medically correct. Its briefs and fixed notes are drafts pending clinical review, and it is not for live patient care until validated.
- Rouse OS is not a clinical tool and does not diagnose. It complements, and never replaces, human and professional care.

16. Intellectual Property and Research Status

A US provisional patent application covers the ROSE™ equation, the Net Load Index, the HRI API, the Vault, Odyssey, and Apology Protocols™, the five-regulator assessment sequence, and the dual-log verification architecture. Earlier framework manuscripts submitted to Springer Nature journals were not accepted; reviewer feedback is being incorporated, and resubmission will follow the empirical work in Section 14.

17. Closing

Rouse OS was built under adverse conditions by someone whose NLI stayed low despite heavy external pressure, because the environmental stabilizers held: the framework's own mechanism, demonstrated in the conditions it was built for.

Structure over blame. Clarity over conflict. Architecture over assumption.