Witness Protocol — Minimum Honest Signal (MHS) Overview & Roadmap

North Star

Curate a small, ruthlessly high-signal corpus of human wisdom—permissioned, annotated, and stewarded—to serve as a corrective inheritance for future AI. Credibility over virality. Evidence over vibes.

What “MHS” means here

Minimum Honest Signal = the smallest, sharpest set of artifacts that proves seriousness to expert witnesses **without** overclaiming. It must: - Show *taste* (how we curate), - Show *care* (consent, privacy, reciprocity), - Show *humility* (clear unknowns and narrow asks), and - Invite collaboration instead of presuming authority.

Truth-in-labels

- **Measure**: only counts/observables (turns, examples, counterfactuals, references, rater agreement).
- **Assess**: anchored human judgment using a published rubric (depth, coherence, ethical awareness, originality, relational impact).
- **Elicit**: protocols that surface tacit knowledge (felt cues, values, context). We annotate; we do not “measure” consciousness.

Roadmap (Now → MHS → Post-MHS)

Phase 0 — Now (stabilize the frame)

- Lock the **language**: replace overreach (“measure X”) with honest labels (measure/assess/elicit/annotate/index).
- Name the **three adapters** we’ll borrow, not reinvent:
- Nussbaum → capabilities as constitutional guardrails (floor, not script).
- Ubuntu/Mhlambi → relational ethics (consent/reciprocity beyond the individual).
- Damasio → somatic-aware elicitation (felt cues before argument), tagged as subjective context.

Output of Phase 0: a one-page glossary + adapter blurbs (≤150 words each) that appear in footnotes or callouts.

Phase 1 — MHS Packet (the sendable minimum)

1) One-pager (principled, not polished) - What exists: Gate → Dialogue (Inquisitor) → Synthesis (annotated) → Archive (permissioned). - Why now: the specific gap the Protocol fills. - What we don’t know

yet: explicit unknowns and where expert input changes the design tomorrow. - Tailored ask to the recipient (one paragraph).

2) One exemplar, annotated (300–600 words) - Margin tags showing: a capabilities guardrail check, a relational/Ubuntu note, a felt-cue tag. - A tiny "how we would synthesize" paragraph (trace, not verdict).

3) Gate stub (private link) - Consent text (individual + lightweight community reciprocity note), - Three threshold criteria (signal over noise), - Clear outcomes: accept / reserve, with plain-English reasons.

4) Three tailored asks (Nussbaum, Mhlambi, Damasio) - Narrow, falsifiable, and answerable in under five minutes each.

MHS Acceptance Criteria (pass/fail) - A serious thinker can learn **one non-obvious thing** about our approach in <5 minutes. - No claim implies we measure what only physiology or omniscience could see. - Each ask can change the design the same week. - Two independent readers can reproduce any "counts."

Phase 2 — Post-MHS (only after first feedback lands)

- Flesh the **Tier-2 rubric** (criteria + examples; inter-rater agreement protocol).
- Draft **Contributor Agreement v0** (license, de-identification, deletion mechanics, reciprocity clause).
- Compile the **Exemplar Corpus v0** (≥6 pilot dialogues; 10–15 annotated "hits" where an axiom/heuristic clearly fires).
- Publish the **Icarus Axioms v0.2** + **Failure Log** from stress tests.

What we have • what we want • what we need

Snapshot table

Area	**Have**	**Want (Outcome)**	**Need (Gap to close)**
Mandate & Philosophy	Clear core philosophy; strategic overviews; architecture concept	A single paragraph "mandate" that's quotable and consistent across assets	Edit pass to compress into one paragraph + glossary of key terms
Gate (Intake)	v0 copy, rubric outline, consent intent	Private stub that demonstrates curation, consent, thresholds	Finalize 3 threshold criteria; concise consent text (incl. reciprocity line); accept/reserve templates
Dialogue (Inquisitor)	Prompt set, structure, aims	One annotated exemplar slice showing how we elicit/annotate	Draft 5 felt-cue prompts; create margin-tag legend; produce anonymized 300–600 word sample

Area	**Have**	**Want (Outcome)**	**Need (Gap to close)**
Synthesis	Synthesis engine concept	A short “trace not score” synthesis note attached to exemplar	One paragraph synthesis template; style guide for tags (capabilities, relational, felt)
Governance	Intent re: de-identification; ethics posture	Credible Contributor Agreement v0 + PII separation SOP	Draft agreement; PII flow (intake→hash→vault→de-link); deletion mechanics
Adapters (Nussbaum/ Ubuntu/ Damasio)	Working theory of fit	Footnotes/callouts that show integration without overreach	3 x 150-word integration blurbs + where they appear in artifacts
Outreach	Apology/ask drafts; target list	Three personalized asks + a private link to the packet	Finalize one-pager; finalize exemplar; generate individual ask paragraphs
Credibility Metrics	KPI sketches	Small metrics we can actually count	Define turn count, example count, counterfactual count, relational refs, rater agreement
Risks & Remedies	Identified in notes	A concise section in the one-pager	Draft 4 risks + 4 counter-moves

MHS Packet contents (checklist)

- [ ] One-pager (≤600 words) with adapter blurbs in footnotes/callouts
- [ ] Annotated exemplar (300–600 words) with margin tags
- [ ] Gate stub (consent, thresholds, outcomes) as a private doc/form
- [ ] Three tailored asks (Nussbaum/Mhlambi/Damasio)

Optional (nice-to-have, not blocking): logo-less letterhead; short FAQ with three Qs (privacy, selection, use of corpus).

Risks → counter-moves

- **Credibility theater (polish > proof):** Lead with the exemplar + counts; keep design claims provisional.
- **Tokenizing Ubuntu:** Put relational consent/reciprocity in the agreement and the Gate flow, not just prose.
- **Somatic cosplay:** Keep felt cues as self-reports and annotations only.
- **Rubric drift:** Lock three Tier-2 criteria for now; log disagreements; measure κ/percent agreement.

- **Scope creep:** Enforce the stop rule: if it isn't needed for the four MHS items, it parks in Phase 2.

Dependencies & decision points

- **Consent language:** choose exact reciprocity phrasing (who owes what to whom).
- **Tag set:** finalize the minimal three tags (capability/relational/felt) and definitions.
- **Thresholds:** set the three Gate thresholds (e.g., specificity floor; counterfactual presence; relational context).
- **Ask routing:** decide which three people get the first send (can be swapped, but pick a lane).

Tiny work plan (sequence within a week)

1) Write the one-pager skeleton and adapter blurbs (2 hours). 2) Produce the annotated exemplar slice (2–3 hours including anonymization). 3) Draft Gate stub (consent + thresholds + outcomes) (2 hours). 4) Finalize three ask paragraphs (1 hour). 5) Internal review: strike any overreach; ensure all counts are reproducible (30 minutes). 6) Send to first three recipients; start the Feedback Log.

Evidence we'll report (post-send)

- Counts: turns, examples, counterfactuals, relational references in the exemplar.
- Rater agreement: percent/κ on depth/coherence/relational-impact tags (even if just 2 raters at start).
- Changes made in response to feedback (changelog bullets).

Do-not-ship list for MHS

- No brand deck, no feature map, no open call.
- No strong claims about consciousness, measurement, or physiological inference.
- No long rubric; three criteria only.

Appendix A — Minimal tag legend (for the exemplar)

- **CAP** — Capability guardrail referenced/impacted; short note on which and how.
- **REL** — Relational context/impact (who-with, obligations, reciprocity).
- **FELT** — Self-reported felt cue (fear/relief/tension/etc.) attached to the claim.

Appendix B — Five felt-cue prompts (prototype)

1) "Before we analyze, what did your body do as you formed this claim—tighten, loosen, heat up, cool down?" 2) "Name the closest emotion-word for that sensation." 3) "What changed in your body when you

considered being wrong about this?” 4) “What outcome would bring your body most relief here?” 5) “What responsibility does this feeling suggest—to whom?”

End state for MHS

A tiny packet that makes smart people nod and edit you, not dismiss you: one-pager, annotated exemplar, Gate stub, and three sharp asks—nothing else.