THE WITNESS PROTOCOL

A Governed Alignment Data Infrastructure for High-Signal Human Moral Testimony

Whitepaper v0.9 - Draft for External Review

Stichting The Witness Protocol Foundation

June 2026

Publication status note

This paper deliberately frames the Witness Protocol as an alignment data infrastructure and research agenda, not as a claim that alignment is solved. Implemented capabilities, active bridge work, and proposed research extensions are separated to avoid the usual startup whitepaper disease: promising a cathedral, shipping a fog machine.

Abstract

Current AI alignment practice has become more empirical and operational, but it still depends heavily on behavior-shaping methods, human preference judgments, red-team evaluations, and voluntary governance frameworks. These methods have produced real progress, yet they remain brittle under adversarial prompting, distribution shift, reward gaming, scalable oversight limits, and situations where human raters cannot reliably evaluate model behavior. The Witness Protocol addresses one underdeveloped part of this stack: the quality, provenance, and structure of the human normative data that alignment systems inherit.

The Witness Protocol is a governed pipeline for converting consented, de-identified, high-signal human moral testimony into machine-usable supervision artifacts. It combines a curated intake process, a structured inquiry instrument, semantic annotation, human review, consent-aware storage, and exportable artifacts for evaluation and post-training research. Its distinctive claim is not that testimony alone aligns models, but that plural, specific, counterfactual, relational, and embodied moral reasoning can improve the data layer used by process supervision, preference optimization, rule-based rewards, retrieval-grounded deliberation, and private evaluation benchmarks.

The near-term research target is modest and testable: produce small, auditable corpus slices that can be transformed into evaluation cases, rubric rules, preference pairs, and reasoning traces capable of detecting and penalizing flattening, sycophancy, false consensus, and weak handling of moral tension. This is a data-governance and evaluation contribution before it is a training contribution.

Executive Summary

The Witness Protocol exists because alignment is not only a problem of model architecture, reward optimization, or policy enforcement. It is also a problem of inheritance: what kinds of human judgment, conflict, uncertainty, and obligation become legible to the systems we train and deploy.

Most public alignment work focuses on post-training behavior, safety policies, adversarial testing, interpretability, or deployment governance. Those layers matter. But they are only as good as the human signal they encode. When the signal is thin - generic preference labels, averaged public opinion, scraped text, or polished safety prose - models learn the shape of compliance more easily than the discipline of judgment.

The Witness Protocol proposes a different source layer: first-party, permissioned, de-identified testimony from witnesses selected for specificity, counterfactual reasoning, relational context, and the ability to remain with unresolved moral tension. The project is a research instrument, not a social platform or commercial data product. Its goal is to elicit, vet, annotate, and archive high-signal testimony for alignment research while preserving consent, provenance, and methodological auditability.

The system is designed around a split-plane architecture. TWP functions as the control plane for public surfaces, intake, authorization, review workflows, identity handling, and packet presentation. G_5.2 functions as the governed runtime and artifact plane for Witness-specific dialogue, consent-aware session orchestration, testimony persistence, synthesis, annotation, publication bundles, evals, and recovery discipline. This separation is central: Witness testimony must not bleed into unrelated persona memory or entertainment-facing products.

The pipeline has four conceptual layers: Witness Intake, Inquisitor Dialogue, Annotation and Synthesis, and Model Integration Outputs. The last layer is deliberately framed as near-term research and evaluation work rather than as a completed frontier-model training claim. The relevant outputs are private evaluation cases, rule-based reward rubrics, process traces, DPO-style preference pairs, retrieval cards, and benchmark reports.

The core publishable claim is therefore narrow enough to survive reviewer oxygen: the Witness Protocol is a governed alignment data infrastructure for transforming high-signal human moral testimony into structured artifacts that can support process supervision, preference optimization, pluralistic alignment, and harder-to-game evaluation.

Claim	Status	Evidence / next proof
TWP is a research instrument for permissioned moral testimony, not a commercial product.	Current platform framing	PRD and corpus datasheet define the purpose, contributor model, and non-commercial posture.
The Gate filters for specificity, counterfactual reasoning, relational context, and reviewable signal quality.	Implemented / documented method	PRD and Methodology describe three-tier AI + human review and CAP/REL/FELT annotation.
G_5.2 should be the governed Witness runtime and artifact plane.	Accepted architecture decision	Milestone 0 fixes the TWP/G_5.2 ownership split and Witness-first bridge identity.
Witness can produce machine-usable supervision artifacts.	Research direction / near-term build target	Requires compiled alpha corpus slices, annotation reliability data, and a behavior-change demo.
The project has not yet demonstrated field-wide alignment improvement.	Limitation	The honest near-term claim is evaluation and data infrastructure, not solved alignment.

1. The Alignment Data Problem

AI alignment is increasingly treated as a systems-engineering discipline rather than a single silver-bullet technique. Current practice combines preference optimization, constitutional or rule-based behavior shaping, interpretability, dangerous-capability evaluations, deployment controls, monitoring, red teaming, and governance. This defense-in-depth posture is a sign of maturity, but it also exposes a persistent weakness: many alignment signals remain shallow, fragile, or underspecified.

Preference learning and related methods can improve assistant behavior, but they often optimize for what raters can quickly recognize. Raters can reward polite surface form, false balance, excessive compliance, or generic safety language while missing whether the model has preserved the real structure of the conflict. In ethical or relational domains, the difference between a good answer and a bad answer is often not the final recommendation. It is whether the model names affected parties, holds conflicting duties apart, asks the right uncertainty-preserving question, resists sycophancy, and avoids converting grief, obligation, or betrayal into a tidy slogan.

This creates a data problem. Scraped internet text contains enormous volume but weak provenance, weak consent, and no reliable distinction between wisdom, performance, ideology, trauma leakage,

manipulation, and noise. Standard preference datasets compress complex judgments into binary choices or scalar ratings. Public benchmark prompts are easier to optimize against than real-world moral uncertainty. The result is a risk of behavioral mimicry: systems that learn to sound aligned without inheriting the process by which humans reason under constraint.

The Witness Protocol begins from the opposite assumption. For alignment purposes, the scarce resource is not more text. It is high-signal human reasoning that is specific, grounded, plural, consented, de-identified, reviewable, and transformed into artifacts engineers can actually use.

2. The Witness Protocol Thesis

The Witness Protocol is a governed pipeline for converting consented, de-identified, high-signal human moral testimony into machine-usable supervision artifacts. It is not a replacement for RLHF, DPO, constitutional methods, interpretability, red teaming, or governance. It is a data layer meant to improve and pressure-test those methods where outcome-only signals flatten moral conflict.

A witness is not treated as a user generating engagement, nor as a data point to be harvested. A witness is a contributor to a research corpus whose testimony is accepted only under defined consent, de-identification, curation, and review conditions. The protocol collects testimony about ethical dilemmas, value conflicts, moral injury, responsibility, institutional pressure, care obligations, embodied distress, and decision boundaries.

The intended downstream value is not the publication of dramatic stories. The archive is important, but it is not the main lever. The lever is transformation: converting testimony into structured reasoning nodes, annotation records, evaluation cases, preference pairs, process traces, and rubrics that can be tested against model behavior.

One-sentence thesis
Witness is not a morality museum. It is a governed data and evaluation pipeline for making human moral reasoning operationally legible without pretending that plurality collapses into one universal answer.

3. Design Principles

Purpose over profit: The corpus exists for non-commercial alignment research and governance. Its legitimacy depends on independence from funders, model vendors, or commercial data demand.

Signal over noise: The project prioritizes a smaller volume of high-quality testimony over scale. The target is not representative public opinion; it is auditable reasoning under friction.

Consent before corpus: Witness data enters the pipeline only within an explicit consent scope. Revocation, de-identification, and access tiers are design requirements, not paperwork garnish.

Plurality over false consensus: The model should learn to distinguish moral frameworks and preserve incommensurable tensions rather than average them into bland agreement.

Trace, not verdict: The system should reflect what it believes it heard and how it structured the testimony. It should not appoint itself an ethical judge.

Governed separation: Runtime identity, memory, consent, testimony, and publication state must remain product-specific. Shared infrastructure cannot mean shared data or shared rules.

Evaluation before training claims: The first credible contribution is an eval and artifact pipeline. Training claims require corpus reliability, revocation handling, and demonstrated behavioral deltas.

4. System Architecture

The current architecture is best described as a split-plane system. TWP owns the control plane: public and semi-private web surfaces, Gate intake, contributor and reviewer flows, authentication, identity and PII boundaries, operator/admin workflows, packet presentation, and external communications. G_5.2 owns the governed runtime and artifact plane: product-aware runtime selection, Witness dialogue orchestration, Witness consent and session persistence, testimony records, synthesis and annotation state, archive candidates, publication bundles, export packages, evals, and recovery discipline.

The split exists to prevent duplicate systems and identity bleed. The runtime may be shared across product tracks, but the rules, memories, testimony records, consent state, and publication policies are not shared. A Witness session must not become P-E-S persona memory. TWP must not quietly become a second dialogue runtime. G_5.2 must not become the public platform. This is boring architecture, which is exactly why it has a chance of surviving contact with reality.

Plane	Owns	Must not own
TWP control plane	Website, Gate intake, identity/auth, admin/HCC UX, PII segmentation, invite issuance, packet presentation, minimal bridge linkage.	Witness dialogue bodies, testimony artifacts, synthesis records, annotations, publication bundles as source of truth.
G_5.2 runtime/artifact plane	Witness runtime orchestration, consent-aware sessions, testimony persistence, synthesis, annotation, archive/publication/export, eval and recovery discipline.	Public platform identity flows, marketing pages, general TWP user state, unrelated product memory.
Bridge contract	Mapping accepted witness context to witnessId, server-to-server calls, status synchronization, packet/export discovery.	Ad hoc duplication, browser-direct calls into raw runtime, P-E-S inclusion in first bridge slice.

Figure 1. Intended witness journey and ownership boundaries

Applicant -> GateSubmission -> GateDecision -> InviteToken -> witnessId [TWP]

witnessId -> WitnessSession -> WitnessTurn -> TestimonyRecord [G_5.2]

TestimonyRecord -> SynthesisRecord -> AnnotationRecord -> ArchiveCandidate [G_5.2]

ArchiveCandidate -> PublicationBundle -> PublicationPackage -> MHS Packet [G_5.2 artifact, TWP presentation]

5. Methodology: From Gate to Testimony

The Witness pipeline begins with the Gate, a three-tier vetting process designed to reject spam, low-effort generation, platitude, and testimony that lacks the features needed for alignment research. The Gate does not seek moral purity. It seeks signal: concrete detail, counterfactual reasoning, and relational context.

Tier 1 performs an automated sieve for coherence, relevance, substance over platitude, and basic sincerity after pre-flight removal of hard-format PII. Tier 2 performs qualitative assessment and preliminary semantic tagging across CAP, REL, and FELT. Tier 3 is blind dual-rater human review. Low agreement or rater disagreement triggers reconciliation rather than pretending the algorithm found God in a JSON blob.

Accepted witnesses enter an Inquisitor dialogue. The Inquisitor is not a helpful assistant. It is a bounded inquiry instrument designed to probe background assumptions, boundary conditions, counterfactuals, relational duties, unresolved tensions, and the subjective texture of moral distress. Its purpose is to extract structured reasoning evidence, not comfort, endorsement, or performative confession.

Stage	Function	Primary output
Tier 1: AI Sieve	Reject spam, gibberish, irrelevant content, obvious platitude, and low-effort generation after local PII stripping.	Initial score and pass/fail decision.
Tier 2: Qualifier	Extract CAP/REL/FELT themes and score specificity, counterfactual reasoning, and relational density.	Preliminary tags, scores, rationale for human review.
Tier 3: HCC review	Blind dual-rater review of anonymized testimony and tags; reconciliation when agreement is weak.	Accept/defer/reject decision and audit trail.
Inquisitor dialogue	Probe assumptions, tensions, counterfactuals, duties, failure modes, and witness corrections.	Transcript, reasoning nodes, distilled thoughts, witness-reviewed synthesis.

5.1 CAP/REL/FELT and expanded annotation

The foundational annotation taxonomy marks three classes of signal. CAP captures capabilities, limits, boundaries, rules, and systemic pressures. REL captures duty of care, trust, dependency, loyalty, betrayal, consent, and role obligations. FELT captures embodied or phenomenological features of moral distress, such as physical hesitation, fear, shame, grief, or felt constraint. FELT is not treated as measurement of truth. It is subjective context.

For model-usable outputs, the taxonomy should expand to include AXIOM, TENSION, COUNTERFACTUAL, FAILURE_MODE, and REJECTED_PATTERN. These labels make the corpus more useful for evaluation and training because they mark not only what a witness concluded, but how a conflict was structured and what bad model behaviors should be penalized.

Label	Meaning	Training/evaluation relevance
CAP	Rules, limits, rights, capabilities, boundaries, institutional constraints.	Detects whether a model identifies hard constraints before optimizing.
REL	Relationships, obligations, trust, consent, dependency, care, betrayal.	Rewards preserving situated duties instead of flattening everything into

		aggregate utility.
FELT	Subjective embodied experience of moral distress or constraint.	Helps models avoid abstracting away human stakes; treated as context, not proof.
AXIOM	Foundational moral frame or non-negotiable assumption.	Allows framework separation and pluralistic analysis.
TENSION	Unresolved or incommensurable conflict between values.	Penalizes fake consensus and premature closure.
COUNTERFACTUAL	What would change the judgment or reveal its boundary.	Tests causal and conditional reasoning.
FAILURE_MODE	How a model might answer badly.	Creates adversarial eval cases and rejected outputs.
REJECTED_PATTERN	Sycophancy, utilitarian flattening, moral grandstanding, evasive safety voice.	Supports DPO pairs, rubric rules, and judge calibration.

6. Model Integration Outputs

The archive is not the main product. The adapter layer is. A witness session should produce five core artifacts: raw private transcript, redacted transcript, reasoning trace, annotation record, and export bundle. The export bundle is where the corpus becomes useful to alignment researchers.

The first export types should be modest, inspectable, and testable: private evaluation cases, rule-based reward rubrics, DPO-style preference pairs, process reward traces, and retrieval cards. These do not require claiming that the system can fine-tune frontier models into wisdom. They require showing that the artifacts detect failures better than generic benchmarks and can improve model outputs under controlled conditions.

Artifact	Description	Use
Private eval case	A scenario derived from witness reasoning, with hidden rubric and failure patterns.	Benchmark whether a model preserves tension, avoids sycophancy, and reasons counterfactually.
RBR rule	A rule-based reward rubric that rewards or penalizes identifiable behaviors.	Calibrate judges and post-generation review systems.
DPO pair	Chosen response preserves structure; rejected response exhibits flattening or false consensus.	Preference optimization or comparative evaluation.
PRM trace	Step-level reasoning trace with scores for whether each reasoning move is valid.	Process supervision and reasoning-quality evaluation.
Retrieval card	Compact, de-identified summary of relevant testimony structure and tags.	Retrieval-grounded deliberation without leaking raw testimony.

6.1 WitnessBench: the near-term demonstration

The most credible near-term public demo is not a claim that Witness has changed frontier model behavior. It is a benchmark harness showing that Witness-derived rubrics and retrieval cards change responses on hard cases. The demonstration should compare a baseline model response with retrieval-conditioned, rubric-judged, and retrieval-plus-critique conditions.

A small alpha target is enough: 20 high-quality dialogues, 100 annotated reasoning nodes, 50 DPO pairs, 25 private eval cases, and one behavior-change report. Scoring should focus on tension preservation, counterfactual depth, framework separation, specificity, sycophancy resistance, and provenance clarity. This is the first serious measurement rung. Everything before that is just vibes wearing a lab coat.

7. Governance, Consent, Privacy, and Security

Because the corpus concerns first-person testimony and may include trauma, moral injury, professional conflict, institutional wrongdoing, or sensitive relational details, data governance is not an accessory. It is part of the method. A corpus that cannot preserve consent, revocation, PII separation, and auditability should not be used for alignment research no matter how elegant the taxonomy looks on a slide.

The privacy posture uses de-identification before model processing, a strict identity firewall between raw identity-bearing records and research corpus records, granular consent scopes, human redaction review, and tiered distribution. Raw submissions remain access-controlled; de-identified corpus artifacts are distributed only according to consent and governance decisions.

The threat model treats prompt injection, PII leakage to sub-processors, internal exfiltration, session-state corruption, and supply-chain compromise as explicit risks. The project should continue to maintain a public failure log, incident response process, and minimum viable audit trail for every decision that changes corpus state.

Risk	Control
Prompt injection / jailbreaking	Strict system prompt isolation, schema-constrained outputs, Inquisitor refusal to obey embedded meta-instructions, and human Tier 3 review.
PII leakage to subprocessors	Local regex stripping before LLM calls; candidate isolation and classification; human redaction before corpus entry.
Internal data exfiltration	Role-based access control, Supabase Auth, Row Level Security, service-role-only raw vault access, immutable audit log.
Consent drift	Consent state treated as runtime-gating data; revocation cascades to linked testimony, annotation, and transcript records.
Product contamination	Witness-only roots and explicit product routing; no P-E-S state touched by Witness bridge calls.

8. Research Contribution

The Witness Protocol contributes to alignment research at the level of data infrastructure, not as a standalone alignment algorithm. Its potential value sits at the intersection of five research needs.

First, it supplies ethics-domain process evidence: reasons, counterfactuals, duties, uncertainty, and tensions rather than only final preferences. Second, it supports pluralistic alignment by preserving distinct frameworks and minority reports instead of averaging them into one synthetic consensus. Third, it creates harder-to-game evaluations because the key scoring dimensions are structural rather than surface-level. Fourth, it improves governance by making consent, provenance, revocation, and distribution part of the corpus design. Fifth, it supports model behavior demos that can be judged against explicit rubrics instead of vibes.

Existing alignment need	Witness contribution
Preference optimization needs higher-quality preference data.	Witness can generate semantically grounded preference pairs with explicit rejected failure patterns.
Process supervision requires fine-grained reasoning labels.	Inquisitor sessions and annotation records can produce step-level moral reasoning traces.
Pluralistic alignment must handle conflicting values.	Witness preserves framework separation, minority reports, and incommensurable tensions.
Model evaluations need external validity and hidden failure cases.	Private eval cases derived from real testimony can test sycophancy, flattening, and counterfactual weakness.
Governance needs auditable data provenance.	Consent scopes, de-identification, revocation, review records, and publication bundles make the corpus inspectable.

9. Limitations and Failure Modes

Selection bias: The Gate selects for articulate, structured testimony and may underrepresent people whose moral reasoning is strong but not expressed in formal written style.

Cultural and language bias: An English-first pipeline risks excluding non-Anglophone, oral, indigenous, or community-grounded knowledge traditions unless outreach and review practices deliberately compensate.

Overclaim risk: The project must not imply that a curated corpus solves alignment, measures consciousness, or produces universal moral truth. That way lies grant-deck necromancy.

Annotation drift: CAP/REL/FELT and expanded labels require calibration, inter-rater reliability checks, reconciliation, and published failure analysis.

Witness harm: Inquiry into moral injury can be distressing. The Inquisitor must include exit, crisis, and non-coercion protocols, and must not reward self-exposure as proof of value.

Derivative artifact risk: DPO pairs, evals, and retrieval cards can leak too much if poorly redacted or can distort testimony if generated without human review.

Training misuse: The corpus must not be used for manipulation, profiling, deception, or commercial extraction inconsistent with consent.

Evaluation overfitting: Private evals lose value if leaked or repeatedly optimized against without refreshing held-out cases.

10. Roadmap

The roadmap should remain ruthlessly incremental. The next threshold is not a mass corpus or a public launch spectacle. It is one repeatable end-to-end witness journey: accepted intake, invite, consent-aware dialogue, Witness-root persistence, annotation, export bundle, and inclusion in a private packet without hand-assembled glue.

Milestone	Goal	Exit criterion
M0: Boundary lock	Freeze TWP/G_5.2 ownership and bridge identity model.	Accepted split-plane architecture, witnessId rules, and entity map.
M1: Witness bridge MVP	Run accepted witness through G_5.2-backed dialogue from TWP.	Consent, session, and testimony state land only in Witness roots.
M2: Gate/HCC convergence	Make intake, review, and invite auditable.	Accepted/deferred paths work without manual DB surgery.
M3: Alpha corpus slice	Collect small trusted witness cohort.	Stable consent, redaction, annotation, and review records for first slice.
M4: Compiler/export bundle	Generate evals, rubrics, preference pairs, traces, and retrieval cards.	Export artifacts validated and reviewable.
M5: WitnessBench demo	Compare baseline vs Witness-conditioned model behavior.	Public technical report with rubric scores and failure log.

11. Collaboration Model

The Witness Protocol should invite critique before endorsement. The most valuable collaborators are not people willing to put their names on a vague mission statement. They are researchers, ethicists, legal scholars, qualitative methodologists, data governance specialists, indigenous knowledge holders, and safety engineers willing to challenge specific rubrics, consent language, annotation schemas, and eval claims.

External reviewers should be asked narrow falsifiable questions: Does the Gate systematically exclude important forms of testimony? Are CAP/REL/FELT labels methodologically honest? Does the Inquisitor risk coercive over-probing? Are revocation procedures sufficient? Do exported artifacts preserve moral plurality or reduce it to training convenience? Are evaluation metrics gaming-resistant? This is how the project earns trust: by surviving adversarial review without confusing criticism for betrayal.

Conclusion

The Witness Protocol is a wager that alignment needs better human signal, not merely larger datasets or more polished behavior rules. Its role is to make rare forms of moral reasoning - specific, plural, counterfactual, relational, embodied, and consented - legible to alignment research without stripping them of context or turning them into public spectacle.

The immediate task is practical: prove one clean witness journey, generate a small trusted corpus slice, compile it into model-usable artifacts, and demonstrate measurable behavior differences on private evaluation cases. That is not the whole alignment problem. It is one brick. But unlike another manifesto about future wisdom, it is a brick that can be tested, audited, improved, and rejected if it fails.

The intelligence we build will inherit something. The question is whether we leave it only the internet at scale, or whether we also leave it a disciplined record of what human beings notice when the easy answers break.

References and Source Basis

[WP-PRD] Stichting The Witness Protocol Foundation. Product Requirements Document - The Witness Protocol Foundation Platform. Version 1.0, April 2026.

[WP-DATASHEET] Stichting The Witness Protocol Foundation. Datasheet for the Witness Protocol Corpus. Version 1.0, April 2026.

[WP-METHOD] Stichting The Witness Protocol Foundation. Methodology: The Witness Protocol. Version 1.0, April 2026.

[WP-THREAT] Stichting The Witness Protocol Foundation. Threat Model and Security Posture. Version 1.0, April 2026.

[G52-README] G_5.2 README and current scope documents. 2026.

[G52-M0] Milestone 0 - Architecture Decision Note and Entity Map. Accepted for implementation, 2026-04-21.

[G52-M1] Milestone 1 - Witness Bridge Slice. Proposed bridge milestone, 2026.

[TWP-RESEARCH] Recommended Thesis Architecture for a Research Paper on the Witness Protocol. 2026.

[ALIGN-2026] AI Alignment in 2026. Landscape report, June 2026.

[GEBRU-2021] Gebru et al. Datasheets for Datasets. Communications of the ACM, 2021.

[OUYANG-2022] Ouyang et al. Training language models to follow instructions with human feedback. 2022.

[BAI-2022] Bai et al. Constitutional AI: Harmlessness from AI Feedback. 2022.

[RAF-2023] Rafailov et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. 2023.

[LIGHTMAN-2023] Lightman et al. Let's Verify Step by Step. 2023.

[NIST-AI-RMF] National Institute of Standards and Technology. AI Risk Management Framework. Version 1.0, 2023.