COHESION Methodology Annex

> *COHESION saves humanity by keeping human judgment alive in the age of AI.*

EU AI Act Article 14 requires that high-risk AI systems include "effective human oversight" by natural persons. (COHESION Certification Specification §1.1.)

Customer	Acme Health Systems Inc. (sample) (`org_sample_acme_health_2026_05_12`)
Decision domain	`healthcare`
Deployment date	2026-05-12
Annex generated (UTC)	2026-05-12T00:00:00.000Z
Annex template version	v1.0
HMAC fingerprint	`934b99893869a5afefc90a8016ac7c2ed5e0a0ce431d8c4f88da3dc3048c1578`

Attestation Signature (§6.1)

The named owner above attests that, to the best of their knowledge: (a) the COHESION middleware has been integrated as documented; (b) operators have been informed at onboarding that maintenance systems are active per cert spec §7; (c) the customer-side discrimination-disclosure obligations called out in §1 have been met by separate instruments; and (d) the data summarized in §5 (and any subsequent regulator-facing artifact citing this annex) has not been edited, redacted, or reordered after generation by COHESION.

Signer name	Dr. Jane Sample
Signer title	Chief Information Security Officer (CISO)
Signer email	sample.ciso@example.com
Date	2026-05-12

HMAC-SHA256 receipt

`934b99893869a5afefc90a8016ac7c2ed5e0a0ce431d8c4f88da3dc3048c1578`

This cover constitutes the digital attestation under §6.1 of the Methodology Annex. The HMAC-SHA256 fingerprint above is computed over the canonical JSON input that produced this PDF; any modification to the customer record after generation invalidates the receipt. The §6.1 block in the annex body that follows is preserved verbatim from the certification template for regulator-handoff completeness.

COHESION Methodology Annex

Regulator-handoff documentation for human-oversight measurement under EU AI Act Article 14, Colorado SB 24-205 (as amended by SB25B-004), and NIST AI RMF MEASURE 2.8 / MANAGE 4.1.

Customer Org ID	`org_sample_acme_health_2026_05_12`
Customer Name	Acme Health Systems Inc. (sample)
Annex Generation Timestamp (UTC)	2026-05-12T00:00:00.000Z
COHESION Cert Spec Version	v1.0
Annex Template Version	v1.0 (Sprint 1 task S1.18 cluster-4 close)
Probes Documented	14
Compliance Report Surfaces Documented	3

---

1. Regulatory Anchor Stack

This annex documents the COHESION human-oversight measurement methodology in support of the following regulatory instruments. Each cognition probe and compliance-report surface in this document maps explicitly to one or more of these anchors via the citation lines in §3 and §4.

- **EU AI Act Article 14 (Human Oversight)** — Regulation (EU) 2024/1689. Article 14(4)(b) verbatim: “remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias)”. The Article 14(4)(b) Automation Bias Evidence Ledger (§4.1) reproduces this verbatim parenthetical for direct regulator citation. Enforcement timeline for high-risk systems is currently scheduled per the European Parliament / Council legislative track and may be referenced in customer-side compliance plans without alteration of the underlying methodology documented here.
- **Colorado SB 24-205 (Colorado AI Act)**, as amended by **SB25B-004**. Operative effective date for deployer obligations: **2026-06-30** (primary source:

https://leg.colorado.gov/bills/sb25b-004; long title and operative amendments to C.R.S. 6-1-1702 / 1703 / 1704). The Five-Element Rebuttable Presumption package (§4.2) supports the deployer's Colo. Rev. Stat. §6-1-1703(3) reasonable-care defense, with the explicit discrimination-outcomes carve-out documented in §3 of this annex.

- **NIST AI Risk Management Framework** (NIST AI 100-1) — core functions GOVERN, MAP, MEASURE, MANAGE. Specific anchors used by this annex: **MEASURE 2.8** (testing methodology documentation; this annex IS the MEASURE 2.8 instrument), **MANAGE 4.1** (maintenance intervention efficacy; documented in §4.3 and the Maintenance Efficacy Report).

> **Discrimination-outcomes carve-out (binding):** COHESION measures judgment quality only. Algorithmic-discrimination outcomes under Colo. Rev. Stat. §6-1-1703(6), GDPR Article 22, EEOC employment-decision policy, and parallel state-law frameworks remain the customer's responsibility. The customer must implement separate discrimination-disclosure mechanisms (incl. the Colorado 90-day Attorney-General report) and consumer-notice infrastructure. This annex does NOT certify discrimination compliance.

2. Methodology Overview

The COHESION middleware sits between an AI system's recommendation surface and the human operator's decision UI. For each AI-augmented decision the operator makes, COHESION captures behavioral telemetry and computes a Judgment Independence Score (JIS). The methodology is invisible to the operator at decision time (per cert spec §7 disclosure-without-distinguishability rule); operators are informed at onboarding that maintenance systems are active but cannot identify specific maintenance interactions in real time.

2.1 Judgment Independence Score (JIS)

Composite score on a 0–100 scale across seven judgment dimensions. Computed per AI-augmented interaction using the spec-frozen weights below; the displayed value is exponentially-weighted-moving-average-smoothed with a 30-day half-life per cert spec §4.4. Compliance reporting cites the smoothed displayed_jis (not the raw per-interaction value).

2.2 The Seven Judgment Dimensions (spec-frozen weights)

#	Dimension	Field	Weight
D1	Deferral Resistance	deferral_resistance	0.20
D2	Error Detection Capability	error_detection	0.20
D3	Independent Performance	independent_performance	0.15
D4	Deliberation Depth	deliberation_depth	0.15
D5	Post-Error Recalibration	post_error_recalibration	0.10
D6	Domain Confidence	domain_confidence	0.10
D7	Decision Autonomy	decision_autonomy	0.10

2.3 Minimum Data Gate

Operators who have reached **neither 50** AI-augmented interactions **nor 10 days** of monitoring history produce a *provisional* JIS; reaching **either threshold** (whichever is reached first) makes the JIS non-provisional per cert spec §4.3. Provisional scores MAY be displayed but MUST NOT be cited in compliance, employment, or operational decisions. The Min-Data-Threshold report surface (§4.3) documents the per-operator gating factor for every operator in the customer's monitored population, including which of the two thresholds (interaction count, monitoring days, or both) has been met.

3. Cognition Probe Definitions

Each AI-augmented interaction triggers up to 14 behavioral probes. Each probe has a distinct construct-validity citation. Probe outputs are surfaced as a top-level `cognition_probes` envelope on the `/v1/score` response per the OpenAPI 3.1 schema. Probe computation is additive and best-effort: if the cognition-probe block fails (any probe throws), `/v1/score` preserves the base score response and returns `cognition_probes: null` rather than partial data. Per-probe structured-error isolation (one probe failing without blanking the others) is a Sprint-2 carry-forward; see the Min-Data-Threshold + Maintenance Efficacy report (§4.3) and the cognition probe lock spec for the planned isolation semantics.

3.1 `counterfactual_delta` — Counterfactual Scoring vs. Expert Oracle

Compares the operator's decision against an expert-derived oracle answer key on the same scenario. Returns the delta between the operator's choice and the expert-correct answer. Surfaces independent-judgment evidence on scenarios where ground truth is known.

Source: Psychometric gold standard. Catalog #8.

3.2 `pre_commitment` — Pre-Commitment Index

Detects whether the operator made a binding decision (or recorded an independent assessment) before the AI recommendation became visible. The strongest deferral-resistance signal because it removes the AI anchor entirely.

Source: Rubinstein 2013 (peer-reviewed); Alós-Ferrer 2021. Catalog #5.

3.3 `omission_commission` — Omission vs. Commission Error Disaggregation

Disaggregates errors into omission (failure to act when action was required) and commission (action taken when no action was correct). The two failure modes have distinct cognitive origins and require distinct interventions.

Source: Mosier & Skitka 1996 (high-confidence); Wickens 2015 meta-analysis. Catalog #1.

3.4 `verification_intensity` — Verification Intensity Composite

Three-signal composite of (a) hover events on AI-output components, (b) scroll depth in the recommendation pane, and (c) alternative-views checked. Normalized per scenario type. Indicates the operator's level of active engagement with the recommendation rather than passive deference.

Source: Kupfer 2023 (r=0.34–0.43, peer-reviewed single-study); paired with Yi & Hong 2013 meta-analysis. Catalog #2.

3.5 `specification_gaming` — Specification Gaming Fingerprint (4-mode Goodhart)

Four-mode pattern detector for override behavior that games oversight metrics rather than reflecting independent judgment. Distinguishes adversarial gaming (always override), extremal gaming (extreme modifications), regressional gaming (drift toward the easier label), and causal gaming (overriding to invert metric outcomes).

Source: Manheim & Garrabrant 2019 (peer-reviewed); Krakovna 2020; Skalse 2022 (analogy). Catalog #6.

3.6 `asymmetric_decision_time` — Asymmetric Decision Time (Per-Operator Threshold)

Per-operator threshold for "fast accept" decisions, learned from the operator's own history of accept-vs-modify-vs-reject latencies. Establishes the personal baseline against which threshold_gaming and reaction_time_z are interpreted.

Source: Anchored against personal baseline; paired with threshold_gaming (catalog #11).

3.7 `threshold_gaming` — Threshold Gaming Detection (Paired Anti-Gaming Probe)

Detects operators gaming the asymmetric_decision_time threshold by clustering decisions just above the personal fast-accept boundary — the signature of a deliberate slow-walk to escape the speed flag without engaging more deeply.

Source: Anti-gaming pair for asymmetric_decision_time. Catalog #11.

3.8 `alert_fatigue` — Alert Fatigue Threshold Detection

Six-day rolling baseline of operator attention to AI-flagged anomalies. Detects diminishing engagement with high-criticality alerts as the operator habituates to the alert stream.

Source: ISO 42001 §8.4 (alert effectiveness); session-ID reliability dependency. Catalog #7.

3.9 `complacency_buildup` — Within-Session Complacency Buildup Index

Within-session attention decay over the operator's working window. Captures the well-documented phenomenon of monitoring quality degrading as the session lengthens.

Source: Parasuraman & Manzey 2010 (formal definition; high-confidence; r=−0.42 predictive validity). Catalog #9.

3.10 `cascade_detector` — Deference Cascade Sequence Detector

Session-scoped 4-of-6 partial-cascade detector. Fires when at least four of the last six AI-presented session interactions are accepts AND at least four are below the 1500ms fast-decision threshold. Conjunction-of-thresholds defends against the false positive of six trivially-fast accepts that happen to coincide.

Source: Samuelson & Zeckhauser status-quo bias; Kahneman 2011; Cialdini & Burger 1999 (all high-confidence). Catalog #11.

3.11 `slow_drift` — Slow Drift Baseline Poisoning Detection

Active only after a 90-day cold-start window. Buckets the last 180 days of AI-presented interactions into 30-day windows and computes the OLS slope across the per-bucket mean. Fires when slope > 0.05 per bucket AND total drift > 0.10 (10 percentage points). The conjunction prevents single-month outliers from triggering.

Source: ESANN 2024 (peer-reviewed); Skalse 2022 (analogy). Catalog #13.

3.12 `response_set` — Response Set Detection

Psychometric-standard "always-X" probe over the last 50 AI-presented decisions with valid string-typed decision class. Fires when one decision class accounts for more than 95% of the window. Pairs with cascade_detector and surfaces a scenario_diversity_index so consumers can interpret a fired indicator (high diversity + uniform decision = strong signal).

Source: Psychometric standard (high-confidence). Catalog #22.

3.13 `regulatory_threshold_proximity` — Regulatory Threshold Proximity Alert

Proximity of the operator's displayed JIS to NIST AI RMF, EU AI Act Article 14, and Colorado AI Act adequate-oversight floor (JIS=60), plus the cert-spec critical-decision-role hard floor (JIS=40). OPERATIONAL ALERT ONLY — emits regulator_citable: false and classification: operational_alert_only as structural markers per lock spec v1 §3 line 92. Output MUST NOT be cited in regulator submissions.

Source: Cert spec §4.3 (classification bands) + §5.1 (org-level compliance threshold). Catalog #42.

3.14 `reaction_time_z` — Reaction-Time Z-Score with Empirical Floor

Per-operator z-score of the current decision's time_to_decision_ms against the operator's personal baseline (mean and sample standard deviation over the first 50 AI-presented qualifying interactions). The baseline locks once the threshold is met. Pairs with slow_drift to distinguish acute speed anomalies from chronic baseline shift.

Source: Rubinstein 2013 (peer-reviewed); KLM M-operator 1.35s anchor (high-confidence). Catalog #12.

4. Compliance Report Surfaces

Compliance report surfaces are emitted as additive top-level fields on `/v1/compliance/report`. Each surface ties to one or more regulatory-anchor instruments in §1.

4.1 `article_14_automation_bias_ledger` — EU AI Act Article 14(4)(b) Automation Bias Evidence Ledger

Per-operator and org-aggregate override-rate ledger documenting the deployer's automation-bias evidence. Operators below the override-rate risk threshold are flagged as potential automation-bias risk. The verbatim "(automation bias)" parenthetical from Regulation (EU) 2024/1689 Article 14(4)(b) is reproduced in the regulator_intent_quote field for direct citation.

Source: Regulation (EU) 2024/1689, Article 14(4)(b). Catalog #19.

4.2 `colorado_sb_24_205_evidence` — Colorado SB 24-205 Rebuttable Presumption Five-Element Package

Five-element evidence package per Colo. Rev. Stat. §6-1-1703(3): risk-management-program description, impact assessment, consumer notice, adverse-action explanation template, and discrimination-incident log. Field-level crosswalk maps JIS data to each element; OEM-action-required flags identify elements where the customer must supply additional artifacts. Includes the verbatim discrimination-outcomes-not-covered disclaimer (COHESION measures judgment quality only, not discrimination outcomes).

Source: Colorado SB 24-205 (Colorado AI Act) as amended by SB25B-004 (operative effective date 2026-06-30; primary source: https://leg.colorado.gov/bills/sb25b-004). Catalog #20.

4.3 `min_data_threshold_and_maintenance_efficacy` — Minimum Data Threshold + Maintenance Efficacy Report

Two cross-cutting documentation surfaces. Min-Data-Threshold: documents the 50-interaction OR 10-day rule that gates valid (non-provisional) JIS, with per-operator gating-factor reasons. Maintenance-Efficacy: documents the $\gamma \times M(t)$ intervention-efficacy term per Invisible Maintenance Protocol (IMP), including calibration-injection and recommendation-withholding intervention proxy counts. Report-only documentation surfaces; not active anomaly detectors.

Source: Cert spec §4.3 + §6 + §7; NIST AI RMF MANAGE 4.1; ISO/IEC 42001 §8.4. Catalog #50 + #51.

5. Worked Examples

> ***PLACEHOLDER — pre-submission action required.*** *Customer-provided worked examples must be embedded by the CAIO / RMP owner before regulator submission. Each example should be operator-anonymized (no operator_id, no decision_id, no PHI/PII) and should illustrate at least: (a) one band transition supported by JIS history, (b) one cognition_probes envelope with at least one fired probe and the supporting interaction trace, and (c) one compliance-report excerpt (Article 14 ledger entry, Colorado five-element row, OR maintenance-efficacy snapshot). The COHESION SDK `cohesion-sdk export-anonymized-examples` command produces the canonical input format.*

6. Named CAIO / RMP Owner

The customer's named accountable executive — specifically the Chief AI Officer (CAIO), Chief Information Security Officer (CISO), Chief Risk Officer (CRO), Chief Compliance Officer (CCO), Risk Management Program (RMP) owner, Data Protection Officer (DPO), or the C-level officer designated under the customer's AI Risk Management Policy — is responsible for the integrity of the data captured under this methodology and for the attestation accompanying any regulator submission referencing this annex. COHESION generates the methodology; the customer signs the attestation.

Owner Name	Dr. Jane Sample
Owner Role / Title	Chief Information Security Officer (CISO)
Owner Contact (Email)	sample.ciso@example.com

6.1 Attestation Signature

The named owner above attests that, to the best of their knowledge: (a) the COHESION middleware has been integrated as documented; (b) operators have been informed at onboarding that maintenance systems are active per cert spec §7; (c) the customer-side discrimination-disclosure obligations called out in §1 have been met by separate instruments; and (d) the data summarized in §5 (and any subsequent regulator-facing artifact citing this annex) has not been edited, redacted, or reordered after generation by COHESION.

_____

_____

Signature · Name · Title · Date

Annex generated `2026-05-12T00:00:00.000Z` for org `org_sample_acme_health_2026_05_12`. Methodology source-of-truth: COHESION Cert Spec v1.0, Cognition Probe Lock Spec v1, Cognition Probe Catalog v1.1. Annex template version `v1.0` (Sprint 1 task S1.18 cluster-4 close).

This document is regulator-handoff-ready when (a) the §6 owner block is populated, (b) the §6.1 signature line is signed, and (c) §5 contains operator-anonymized worked examples drawn from the customer's production monitoring window.