The Human Layer Audit: Measuring Accountability in AI Systems

Ahmad Noureddine

Timer · ORCID 0009-0003-1669-4553

2026-04-07 · DOI 10.5281/zenodo.19453025

The first paper in this series established why the Human Layer matters: AI systems designed to amplify human judgment outperform those designed for replacement. [20] The second paper specified how to build it: five architectural components that form a dependency graph, not a checklist. [21] Decision gates, escalation protocols, accountability structures, override mechanisms, and trust calibration interfaces.

This paper addresses the question that follows from both: how do you know if it actually works?

An architecture that cannot be measured cannot be enforced. A specification that cannot be scored cannot be audited. And a human layer that exists on paper but fails under pressure is worse than no human layer at all, because it creates the illusion of oversight while providing none of its function.

Paper 1 introduced a six-question audit as a starting point. Those questions were binary: yes or no. Binary assessment tells you whether a component exists. It does not tell you whether it functions. It does not tell you whether it degrades under load. It does not distinguish between a decision gate that produces real human judgment and one that produces rubber-stamp approvals at three seconds per click.

This paper replaces that binary assessment with a maturity model. Five components. Five levels each. Scored against risk tiers that determine the minimum acceptable standard. The result is a framework that builders can use to design, auditors can use to evaluate, and regulators can reference when defining what "human oversight" must actually mean in practice.

The existing regulatory and standards landscape tells organizations what to do. The EU AI Act mandates human oversight. [3] The NIST AI Risk Management Framework structures governance around four functions. [11] ISO/IEC 42001 requires accountability assignment and operational controls. [9] The OECD AI Principles, adopted by 47 countries and updated in 2024, require that AI actors "ensure traceability, including in relation to datasets, processes and decisions made during the AI system lifecycle." [22] IEEE 7000-2021 provides a methodology for embedding ethical values into system design. [23] The G7 Hiroshima AI Process launched a Voluntary Reporting Framework in February 2025 to enhance transparency in advanced AI development. [24]

None of these frameworks measures whether the human layer is structurally sound. They define obligations. They do not provide a scoring instrument. This paper does.

1. Risk Tiers

Not every AI system requires the same level of human oversight. A content recommendation engine and a criminal sentencing algorithm operate at fundamentally different stakes. Applying the same architectural requirements to both produces either over-engineering in low-risk contexts or dangerous under-specification in high-risk ones.

The EU AI Act already classifies AI systems into risk categories. [3] But its classification tells organizations which regulatory tier they fall into. It does not specify what their human layer architecture must look like at each tier. This framework extends risk classification into architectural requirements.

Three tiers, defined by consequence severity and reversibility:

Tier 1: Advisory. The AI system informs human decisions but does not execute actions independently. Outputs are low-consequence and reversible. Examples include content recommendations, scheduling optimization, document summarization, and search ranking. The human operates with full autonomy. The AI provides input. If the input is wrong, the cost of correction is minimal.

Tier 2: Collaborative. The AI system and human share the decision process. Outputs carry moderate consequence and are partially reversible. Examples include diagnostic triage, credit pre-screening, asset pre-verification, insurance claim assessment, and hiring shortlisting. The AI narrows the decision space. The human makes the consequential judgment within that space. Errors are recoverable but carry measurable cost in time, resources, or stakeholder trust.

Tier 3: Consequential. AI output directly affects individual rights, safety, liberty, or significant financial exposure. Outputs are irreversible or carry high-cost reversal. Examples include criminal risk scoring, autonomous trading execution, medical treatment recommendations, biometric identification for law enforcement, and institutional-grade asset approval for regulated markets. At this tier, the human layer is not an enhancement. It is a legal and structural requirement.

Each tier sets different minimum maturity requirements across the five components defined in Paper 2. Tier 1 systems must achieve at least Level 2 (Functional) across all components. Tier 2 systems require Level 3 (Structural) minimum. Tier 3 systems require Level 4 (Adaptive) minimum, with no component scoring below Level 3.

The tier classification is determined by the highest-consequence output the system can produce, not the average. A diagnostic AI that handles routine screenings at Tier 1 but also flags potential malignancies operates at Tier 2 for the entire system, because the escalation pathway leads to consequential clinical decisions. A tokenized finance platform that automates routine document checks but also approves institutional-grade assets for market operates at Tier 3, because the approval decision is irreversible and carries regulatory exposure.

The conservative principle: if the tier is ambiguous, classify upward.

A system cannot reduce its tier classification by architecturally isolating high-consequence outputs into separately audited subsystems. The tier applies to the integrated system as deployed. Scope narrowing is the oldest audit evasion tactic in regulated industries. This framework does not permit it.

---

2. The Maturity Model

The maturity model evaluates each of the five Human Layer components across five levels. The result is a 25-cell matrix that maps an organization's current state and provides a concrete improvement pathway.

Level 0: Absent. The component does not exist in the system. No decision gate, no escalation protocol, no accountability mapping, no override mechanism, or no trust calibration interface has been designed or implemented. The system operates without this function entirely.

Level 1: Nominal. The component exists in policy, documentation, or stated intention but is not enforced by the system architecture. A decision gate is described in a compliance document but the system does not actually halt execution pending human authorization. An accountability structure names responsible parties in an organizational chart but the system does not log which human authorized which output. The component is decorative. It satisfies a checkbox. It does not produce oversight.

Level 2: Functional. The component is implemented and operational under normal conditions. Decision gates halt execution and require human input. Escalation protocols route uncertain cases. Accountability structures map outputs to humans. Override mechanisms are accessible at the interface. Trust calibration provides reliability signals. But the component is not measured, is not audited, and its effectiveness is not monitored. It works when conditions are favorable. Whether it works under load, under stress, or over time is unknown.

Level 3: Structural. The component is enforced by the system, measured via telemetry, and produces auditable logs. It cannot be bypassed without triggering an exception. Decision gates log the human's action, the time elapsed, and the delta between the system's recommendation and the human's decision. Escalation protocols track routing accuracy and response time. Override events capture justification. Trust calibration telemetry calculates RAIR (Relative Positive AI Reliance) and RSR (Relative Positive Self-Reliance) metrics. [15] The persistence layer required by Paper 2 is fully operational. The component functions as architectural infrastructure, not as a feature that can be toggled off.

Defining RAIR and RSR. These metrics are central to Level 3 and Level 4 differentiation and must be computable from the persistence layer, not merely referenced.

RAIR (Relative Positive AI Reliance) = the number of cases where the human updated an initially incorrect decision based on correct AI advice, divided by the total number of cases where the AI was correct and the human was initially incorrect. A high RAIR means the human trusts the AI when the AI is right.

RSR (Relative Positive Self-Reliance) = the number of cases where the human rejected incorrect AI advice in favor of their own correct judgment, divided by the total number of cases where the AI was incorrect and the human was initially correct. A high RSR means the human trusts themselves when the AI is wrong.

A worked example: a diagnostic triage system processes 1,000 cases in a month. In 80 cases, the AI was correct and the human initially disagreed. Of those, the human updated their decision in 60 cases. RAIR = 60/80 = 0.75. In 40 cases, the AI was incorrect and the human initially had the right answer. The human held their ground in 30 cases. RSR = 30/40 = 0.75. Both metrics at 0.75 indicate reasonable calibration. If RAIR climbs toward 1.0 while RSR drops toward 0.0, the human is deferring to the AI indiscriminately. If RSR climbs while RAIR drops, the human is ignoring the AI entirely. A well-calibrated system maintains both metrics above a domain-specific baseline.

The following alert thresholds are illustrative, not normative, and are subject to the same domain-justification requirement as all other thresholds in this framework. RAIR or RSR sustained below 0.5 over a 90-day window warrants investigation into whether the human layer is engaged. Both metrics simultaneously below 0.3 indicates the human layer is not functioning as calibrated: the human is neither learning from the AI when it is right nor correcting the AI when it is wrong. Conversely, RAIR sustained above 0.95 with RSR below 0.2 signals uncritical deference. These ranges must be calibrated against the specific domain, task complexity, and AI model accuracy. Adopting them without justification does not satisfy the threshold documentation requirement.

Level 4: Adaptive. Everything in Level 3, plus the system monitors its own human layer for degradation and triggers corrective action. Decision gates flag when approval times fall below minimum thresholds, indicating rubber-stamp behavior. Escalation protocols detect when escalation frequency drops to zero in edge-case categories, indicating the system may be suppressing uncertainty signals. Accountability structures alert when a single individual is accumulating decision volume beyond reasonable review capacity. Override monitoring detects when override rates drop to zero (indicating automation bias) or spike without pattern (indicating algorithm aversion). Trust calibration tracks whether RAIR and RSR metrics are diverging over time, signaling miscalibration.

The ground truth problem. RAIR and RSR require knowing which party was correct. In many Tier 3 environments, the correct answer is not known immediately. A criminal risk score may not be validated for years. A long-term medical prognosis may take months to resolve. Level 4 monitoring cannot stall while waiting for outcomes. The architecture must use dual-track calibration.

Track A (immediate): the system uses inter-rater reliability as a process proxy. When the AI and human disagree, the system triggers a blind second opinion from another qualified human. High agreement between independent humans serves as the immediate proxy for ground truth. This does not replace outcome data. It provides a real-time signal that the human layer is producing consistent, reasoned judgments rather than rubber-stamp approvals or arbitrary overrides.

Track B (retrospective): once the actual outcome is known (weeks, months, or years later), the system performs a retrospective true-up. It recalculates RAIR and RSR for that decision cohort using verified outcomes and adjusts current escalation thresholds based on the recalibrated data.

Level 4 systems must document the correlation between their proxy signals and actual outcomes. If the proxy reliably predicts long-term accuracy, the immediate monitoring is valid. If it does not, the proxy methodology must be recalibrated. The persistence layer must retain sufficient data to perform this retrospective analysis for the full duration of the outcome window.

At Level 4, the human layer is self-monitoring. It does not wait for an auditor to discover that oversight has degraded. It detects degradation in real time and escalates. This is the level at which the architecture earns institutional trust, because the system itself is watching whether the human layer is still functioning.

Component-Level Maturity Criteria

The following defines what each level looks like for each component. These are the criteria against which self-assessment and external audit are conducted.

Decision Gates. Level 0: No human checkpoint exists before consequential actions. Level 1: Policy states human approval is required; system does not enforce it. Level 2: System halts at gate; human must act to proceed. Level 3: Gate logs action, timing, and recommendation-vs-decision delta; bypass triggers exception. Level 4: System monitors approval patterns; uses complexity-adjusted engagement baselines rather than fixed minimum times (expected engagement time varies by case difficulty, making gaming harder to predict; complexity classification must be derived from measurable case attributes such as input volume, variable count, regulatory exposure, and precedent availability, not from operator self-assessment; for example, in an asset verification pipeline, a routine document check against a known counterparty is low-complexity while a novel asset class with cross-jurisdictional regulatory exposure is high-complexity, and the expected engagement time must differ accordingly); deploys verification sampling by inserting known-answer cases where the AI recommendation is intentionally incorrect to test whether the human catches the error; flags sustained patterns of approval without substantive modification; alerts supervisor when gate engagement degrades. Probe safety constraint: every verification sampling case must be flagged in metadata at the moment of generation, and the system's execution layer must intercept any approved probe before it reaches production. A failed probe reveals a human engagement problem. It must never produce a real-world consequence. In Tier 3 environments, the intercept logic must be logically separated from the decision logic to ensure that a failure in the primary system cannot disable the safety check. Verification sampling programs must be approved by the system's governance authority before deployment. Operators must be informed that verification sampling is active, though not which specific cases are probes. Sampling frequency must be calibrated to avoid interfering with operational throughput while maintaining statistical validity for engagement assessment.

Escalation Protocols. Level 0: System defaults to model output on all inputs, including edge cases. Level 1: Documentation describes when escalation should occur; system does not detect uncertainty. Level 2: System detects uncertainty and routes to human; escalation triggers are functional. Level 3: Escalation logs track routing, response time, and outcome; every chain terminates in a natural person; thresholds are calibrated against domain data. Level 4: System monitors escalation frequency by category; detects suppressed uncertainty; recalibrates thresholds based on outcome data; alerts when escalation pathways are underused.

Accountability Structures. Level 0: No mapping exists between system outputs and responsible humans. Level 1: Organizational chart names responsible roles; system does not enforce assignment. Level 2: Each consequential output is mapped to a named human at point of decision. Level 3: Pre-assignment is enforced before output; delegation failover is explicit; logs capture identity, authority level, and competency verification. Level 4: System detects decision volume concentration; flags when single individuals exceed review capacity; alerts when accountability gaps emerge from personnel changes or system updates.

Override Mechanisms. Level 0: No mechanism to reject or modify AI output at the decision interface. Level 1: Override exists in admin panel or requires technical intervention. Level 2: Override is accessible at the decision interface without technical access. Level 3: Override includes cognitive forcing functions; requires element-level endorsement rather than blanket approval; logs justification; tested under normal conditions. Level 4: System monitors override frequency; detects automation bias (zero overrides) and algorithm aversion (indiscriminate overrides); tests override under degraded conditions (stress, time pressure, cognitive load); verification sampling confirms that humans exercise overrides when the AI is intentionally wrong, not just when the interface makes it convenient.

Trust Calibration Interfaces. Level 0: No reliability signals provided to the human beyond model output. Level 1: Confidence score displayed; no external calibration signals. Level 2: Historical accuracy rates, boundary conditions, and distribution shift flags provided; reliability signals are external to model self-assessment. Level 3: RAIR and RSR metrics calculated from telemetry; human-AI disagreement rates logged; dependency indicators monitored (human performance degrades when AI is unavailable). Level 4: System detects calibration drift over time; alerts when RAIR/RSR diverge; monitors for dependency formation; adjusts recommendation frequency to preserve human agency.

Threshold Documentation Requirement

The maturity criteria reference thresholds throughout: minimum gate engagement times, maximum decision volumes per human, escalation frequency baselines, override rate bounds. These thresholds are domain-specific and cannot be universalized. A 10-second decision gate is appropriate for a credit pre-screening; it is dangerously fast for an institutional asset approval. A human reviewing 150 content moderation flags per day may be within capacity; a human reviewing 150 sentencing risk scores per day almost certainly is not.

The framework does not prescribe universal numbers. It requires that organizations define, document, and justify their own thresholds as part of audit evidence. The following ranges are illustrative, not normative:

Decision gate minimum engagement time: Tier 1, 3+ seconds. Tier 2, 15+ seconds. Tier 3, domain-defined but documented and justified against task complexity analysis.

Maximum consequential decisions per human per day: Tier 1, organization-defined. Tier 2, documented with fatigue analysis. Tier 3, capacity-limited with mandatory rotation or workload caps.

Escalation frequency baseline: if a Tier 2 or Tier 3 system shows zero escalations over a 30-day period, the threshold calibration must be reviewed. Zero escalation in a system processing edge cases is a signal that the detection mechanism is not functioning, not that every case was clear.

Override rate bounds: a sustained 0% override rate over 30+ days in a Tier 2 or Tier 3 system triggers automation bias review. A sustained override rate above 40% triggers algorithm aversion review and potential model recalibration.

Any threshold chosen must be documented in the system's audit evidence, including the rationale for the chosen value, the data used to calibrate it, and the recalibration schedule. An auditor evaluating the system does not check whether the threshold matches a universal standard. The auditor checks whether the threshold exists, is justified, and is enforced. Adopting the illustrative ranges provided in this paper without domain-specific justification does not satisfy the threshold documentation requirement and constitutes an audit deficiency.

---

3. Identity Anchoring

Identity anchoring is the cross-cutting requirement that accountability structures depend on but that existing frameworks consistently underspecify. It answers the operational question: when the system produces a consequential output, which specific natural person authorized it, and can that person be identified after the fact?

Paper 2 established the principle: every consequential output must map to a named human. [21] It also identified the two failure modes: diffused accountability, where responsibility distributes across teams, committees, and vendors until no one is identifiable; and moral crumple zones, where a human is positioned as oversight but lacks the information, authority, or response time to meaningfully control the outcome. [17]

This section specifies how identity anchoring must be implemented.

Pre-assignment. The accountable human must be identified before the output is produced, not determined after the outcome is known. This means the system must know, at the point of the decision gate, which person is authorized to act. Post-hoc accountability assignment, where an organization determines who was responsible only after a failure, is not accountability. It is blame distribution.

Competency requirement. The assigned human must have domain authority, not just system access. A junior operator with login credentials is not equivalent to a domain specialist with decision-making competence. ISO 42001 Clause 5 requires that accountability be assigned to persons with "necessary competence, training and authority." [9] The EU AI Act Article 26(2) uses identical language. [3] Identity anchoring enforces this by requiring the system to verify competency level before accepting a human's authorization at Tier 2 and Tier 3 gates.

Delegation protocol. When the assigned human is unavailable, the system must follow an explicit delegation chain. Delegation cannot be implicit. The system cannot default to "whoever is logged in." The delegation chain must be pre-defined, documented, and the delegated person must meet the same competency requirement as the primary. If no qualified delegate is available, the system must queue the decision rather than proceed without authorized oversight.

Anti-diffusion rule. Accountability cannot distribute across "the team," "the committee," or "the AI." One name per consequential decision. Committees can advise. Working groups can review. But when the decision gate closes, one natural person's identity is logged as the authorizer. This is the architectural enforcement of what the OECD calls ensuring that "AI actors should be accountable for the proper functioning of AI systems." [22] "AI actors" in this specification means identifiable humans with pre-assigned authority.

Contribution trace. The anti-diffusion rule does not prohibit collaborative decision-making. Medical boards deliberate. Credit committees vote. Engineering teams review. The rule requires that when the gate closes, one person signs. But the persistence layer must allow that primary anchor to tag the collaborators, specialist consultations, and committee votes that informed their decision. This serves two purposes. First, it protects the primary anchor from becoming a moral crumple zone by documenting that their authorization was informed by qualified input, not that they were simply the last signature in a chain. Second, it preserves the evidentiary trail that demonstrates meaningful control: the anchor considered expert input and made a judgment, rather than rubber-stamping a committee recommendation they had no independent basis to evaluate. The primary anchor remains solely accountable. The contribution trace documents how they exercised that accountability.

The AssetLink case. When we designed the asset verification pipeline for our tokenized finance infrastructure, the AI achieved 95%+ accuracy on document validation. The system could have processed institutional-grade assets end-to-end without human review. We built the decision gate anyway, because in regulated finance, a verification error is not a rounding error. It is a compliance failure with counterparty, regulatory, and reputational consequences.

But the decision gate alone was not sufficient. We needed to know which human approved each asset, whether that human had the authority and competency to make the approval, and whether the approval represented substantive review or rubber-stamping. Identity anchoring meant pre-assigning senior compliance reviewers to each asset class, requiring domain-specific certification before granting gate authorization, logging reviewer identity with every approval, and monitoring approval velocity to detect when review times dropped below meaningful engagement thresholds. The human layer worked not because we added a checkbox before the final step. It worked because the identity behind the checkbox was verified, competent, and accountable.

---

4. Compliance Traceability

Paper 2 specified a persistence layer as the substrate on which all five components become measurable. [21] This section defines what that persistence layer must contain and how it maps to regulatory reporting obligations.

The audit trail must capture, at minimum, the following data for every consequential system interaction:

System state: The AI's recommendation, confidence level, and any uncertainty flags at the point of the decision gate.

Human action: The specific action taken: approve, modify, reject, escalate, or override. For modifications and overrides, the specific elements changed and justification provided.

Identity record: The authenticated identity of the human who acted, their pre-assigned authority level, and competency verification status.

Temporal data: Timestamp of the AI recommendation, timestamp of human action, and elapsed time between them.

Delta record: The difference between the AI's recommendation and the human's final action. This is the signal that distinguishes active oversight from passive approval.

Escalation trace: If the interaction triggered an escalation, the routing path, destination, response time, and outcome.

This data must be retained for a period determined by the regulatory environment. The EU AI Act requires that logs be "kept for a period that is appropriate in the light of the intended purpose of the high-risk AI system." [3] ISO 42001 Clause 9 requires "performance evaluation" records that enable ongoing assessment of AI system effectiveness. [9] NIST AI RMF's MEASURE function requires documentation sufficient to "quantify, assess, benchmark, and monitor AI risk." [11]

The compliance traceability layer is not a separate system. It is the persistence layer from Paper 2, structured to serve triple duty: operational telemetry for the builder, audit evidence for the regulator, and trust signal for the stakeholder.

The mapping to major regulatory frameworks:

The EU AI Act Article 14 requires evidence that human oversight functions "as intended." The compliance trace provides that evidence through delta records and temporal data. If approval times cluster at two seconds with zero modifications, the trace documents that the oversight is nominal.

The NIST AI RMF GOVERN function requires "identified stakeholders responsible for security, compliance, and decision-making." [11] The identity record satisfies this. MEASURE requires quantifiable risk monitoring. RAIR and RSR metrics calculated from the trace satisfy this.

ISO 42001 Clause 8 requires operational controls for AI systems. The decision gate and escalation logs document those controls. Clause 9 requires performance evaluation. The temporal and delta data enable it.

The OECD AI Principles require traceability "including in relation to datasets, processes and decisions made during the AI system lifecycle." [22] The compliance trace, taken as a whole, is the implementation of that requirement.

Agent identity in agentic systems. The compliance trace as specified above captures the human side of the accountability chain. In systems where AI agents act with delegated authority (executing trades, performing compliance screening, processing verifications), the persistence layer must also capture the agent side: which agent performed the action, with what credentials, at what trust level, and under what policy constraints. The

accountability chain in an agentic system has two endpoints. The human who authorized and the agent that executed. If either side is unidentified, the chain is broken. This means the persistence layer must log cryptographically verifiable agent identity alongside human identity at every decision point. Agent trust scores, behavioral baselines, anomaly flags, and credential status are the machine-side equivalent of the human-side RAIR/RSR metrics and override telemetry. Singapore's Model AI Governance Framework for Agentic AI already recognizes this requirement: responsibility must be defined across all actors in the agent lifecycle, including the agents themselves. [12] A compliance trace that tracks the human but not the agent that acted on the human's authorization is incomplete.

Telemetry governance. The persistence layer captures high-resolution data about individual human decision-making. This creates a legitimate concern: if humans know their disagreement rates, override patterns, and engagement times are being tracked and scored, they face psychological pressure to conform to the AI's output rather than exercise independent judgment. The telemetry would then produce the very automation bias it is designed to detect.

The governance principle is as follows. RAIR, RSR, override rates, and engagement metrics must be used for system calibration, not individual performance evaluation. The purpose of the telemetry is to assess whether the human layer as an architectural component is functioning, not to rank individual humans against each other or penalize disagreement with the AI. Individual-level data may only be reviewed when a gross negligence threshold is crossed (e.g., a reviewer approves a verification sampling case where the AI was intentionally wrong) or when aggregate patterns indicate a systemic failure requiring investigation. This boundary must be documented in the system's data governance policy and disclosed to all humans operating within the human layer. Undisclosed behavioral surveillance of oversight personnel degrades the trust that the human layer depends on to function.

---

5. Escalation Thresholds

Paper 2 established that escalation protocols must be triggered by system-detected uncertainty, not by human self-assessment. [21] The automation bias literature confirms that humans systematically fail to escalate when escalation is most needed. [5] This section specifies how organizations define, calibrate, and maintain escalation thresholds.

Threshold definition. An escalation threshold is the system-detected condition under which AI output is routed to human judgment rather than proceeding through standard automated or semi-automated processing. Thresholds are domain-specific and risk-tier-dependent. They are not universal constants. A threshold that is appropriate for a content recommendation engine is dangerously permissive for a diagnostic triage system.

Threshold calibration. Thresholds must be set against historical accuracy data for the specific domain and use case. The initial calibration requires a baseline period during which all system outputs are reviewed by qualified humans. The divergence between system recommendations and human judgments during this

baseline establishes the accuracy profile. Escalation thresholds are then set at the point where system accuracy drops below an acceptable level for the risk tier.

For Tier 1 systems, thresholds may allow the system to proceed on outputs where confidence exceeds 90%, escalating only cases below that threshold. For Tier 2 systems, the threshold must be more conservative, and the escalation must route to a domain specialist, not a generalist. For Tier 3 systems, the threshold architecture must account for adversarial conditions, distribution shifts, and novel inputs that fall outside the model's training distribution entirely.

Recalibration cadence. Thresholds are not static. Model performance drifts. Data distributions shift. User populations change. An escalation threshold set during initial deployment may be dangerously miscalibrated six months later. The recalibration cadence must be defined at deployment and enforced by the system, not left to manual review schedules that degrade under operational pressure.

For Tier 1 systems, quarterly recalibration may be sufficient. For Tier 2, monthly review of escalation telemetry is the minimum. For Tier 3, continuous monitoring with automated alerts when escalation metrics deviate from baseline is required.

Threshold verification. How do you know your escalation thresholds are correctly set? This is the "who verifies the verifiers" problem. The answer comes from the persistence layer. RAIR and RSR metrics calculated from decision gate telemetry provide the feedback signal. If RAIR is high (humans are updating their decisions based on AI input) and RSR is also high (humans are correctly rejecting bad AI input), the thresholds are well-calibrated. If RAIR drops (humans are ignoring useful AI input) or RSR drops (humans are accepting bad AI input), the thresholds need adjustment. The telemetry is the verifier.

---

6. Integrated Scoring

The five components and five levels produce a maximum score of 20 (each component scored 0-4). But the scoring methodology must prevent a critical failure mode: averaging away weaknesses.

A system that scores Level 4 on trust calibration but Level 1 on override mechanisms has a decorative override function. Averaging the two produces a respectable-looking mean score that masks a structural deficiency. The Human Layer Audit uses floor-based scoring, not mean-based scoring. The overall assessment is determined by the lowest component score, not the average.

Scoring rules:

The Human Layer Score is the sum of all five component scores (range: 0-20). But the assessment band is capped by the minimum component score. A system with four components at Level 4 and one at Level 1 is assessed as if all five were at Level 1. The chain is as strong as its weakest link. This is the architectural principle from Paper 2: the five components form a dependency graph. Removing or degrading any one breaks the guarantees provided by the others. [21]

Assessment bands:

0-4: No Human Layer. The system operates without meaningful human oversight. One or more components are absent. No regulatory framework currently in force or approaching enforcement would consider this system compliant for high-risk deployment.

5-9: Decorative. Components exist in policy or partial implementation but are not enforced or measured. The human layer provides the appearance of oversight without its function. This is the compliance theater zone. Organizations at this level are at highest risk of the moral crumple zone failure: a human is nominally responsible but structurally unable to exercise meaningful control.

10-14: Operational. Components function under normal conditions but lack adaptive monitoring. The system works when things go right. Whether it works under stress, over time, or in edge cases is unverified. Adequate for Tier 1 systems. Insufficient for Tier 2 or Tier 3 without documented improvement roadmap.

15-17: Structural. Components are enforced, measured, and produce auditable evidence. The human layer is architectural infrastructure. Suitable for Tier 2 deployment. Tier 3 systems at this level should demonstrate active progress toward Level 4 on all components.

18-20: Institutional. Components are structural and adaptive. The system monitors its own human layer for degradation and triggers corrective action. Ready for Tier 3 deployment in regulated markets. This is the level at which the architecture earns institutional trust, because the system does not rely on external auditors to discover that oversight has eroded.

Tier-specific minimums:

Tier 1 systems must achieve a minimum Human Layer Score of 10. No component below Level 2. Tier 2 systems must achieve a minimum score of 15. No component below Level 3. Tier 3 systems must achieve a minimum score of 18. No component below Level 3. At least three components at Level 4.

A Tier 3 system scoring 17 fails the audit regardless of individual component distribution. A Tier 2 system with one component at Level 2 fails regardless of total score. The floor is the standard.

Self-assessment versus independent audit. The self-assessment instrument in Appendix A is a diagnostic tool. It enables builders to identify gaps, prioritize improvements, and track progress. It is not sufficient for regulatory compliance claims or public representations of Human Layer maturity. Tier 1 systems may rely on self-assessment. Tier 2 systems that use their Human Layer Score for regulatory submissions, investor communications, or contractual representations must obtain independent verification from a qualified auditor. Tier 3 systems require independent audit as a condition of a valid score. A self-assessed Tier 3 score is a working document, not a compliance determination.

Auditor qualification. A qualified auditor must demonstrate competency across three domains: AI system architecture (understanding of how the five components are implemented technically, including telemetry design, gate enforcement, and escalation routing), regulatory mapping (working knowledge of at least one major framework among the EU AI Act, NIST AI RMF, or ISO 42001), and operational domain expertise (direct experience in the industry where the system under review is deployed). The auditor must be independent of the organization being audited, meaning no financial relationship, employment, or advisory

role within the preceding 24 months. For Tier 3 systems, the audit team must include at least two individuals collectively covering all three competency domains. This framework does not define a certification body or accreditation pathway. It defines the minimum competency floor that auditors must meet for their assessment to constitute a valid score.

---

7. What This Audit Makes Possible

The Human Layer Audit is not a bureaucratic exercise. It is an engineering tool.

For builders, the maturity model provides a concrete development roadmap. Instead of vague mandates to "ensure human oversight," the framework specifies exactly what Level 3 decision gates look like, what telemetry they produce, and what distinguishes them from Level 2. The improvement pathway from Functional to Structural to Adaptive is defined in measurable terms. An engineering team can scope the work, estimate the effort, and track progress against specific criteria.

For auditors, the scoring methodology provides a standardized assessment instrument. Rather than subjective evaluation of whether a system "has" human oversight, the audit produces a numerical score, a component-level breakdown, and a tier-specific compliance determination. Two auditors evaluating the same system should arrive at substantially similar scores, because the criteria are concrete and the evidence requirements are defined.

For regulators, the framework translates high-level mandates into enforceable specifications. The EU AI Act requires human oversight for high-risk systems but does not define what oversight must measurably look like in a deployed architecture. The NIST AI RMF requires risk management but does not specify how to score the human layer's contribution to that management. The Human Layer Audit provides the measurement layer that these frameworks lack.

For investors, the Human Layer Score is a due diligence signal. A company claiming AI-powered operations in a regulated market can be evaluated on whether its human layer is structural (Level 3+) or decorative (Level 1). The score is verifiable against the compliance trace. The question is no longer "does this company have human oversight?" The question becomes "what is their Human Layer Score, and does it meet the minimum for their risk tier?" The Human Layer Score quantifies architectural maturity. The economic implications of that score, including compliance cost, risk exposure, and the capital incentive structures that drive organizations toward or away from structural oversight, are addressed in the final paper of this series.

The gap in the current landscape is specific and measurable. Organizations have frameworks that tell them what principles to follow (OECD), what processes to implement (IEEE 7000), what risks to manage (NIST AI RMF), and what standards to certify against (ISO 42001). What they lack is a scoring instrument that measures whether the human layer in their deployed AI system is functionally sound, structurally sound, and resilient under the conditions where it matters most.

This paper provides that instrument. The self-assessment tool in Appendix A operationalizes it for immediate use.

Appendix A: The Human Layer Audit: Self-Assessment Instrument

This instrument is designed to be extracted, distributed, and used independently of the paper. Hand it to a CTO, a compliance officer, or a regulator. The scoring methodology is self-contained.

Step 1: Classify Your Risk Tier

Identify the highest-consequence output your AI system can produce.

Does the output directly affect individual rights, safety, liberty, or significant financial exposure with irreversible or high-cost-reversal consequences? → Tier 3: Consequential.

Does the output carry moderate consequence with partial reversibility, where the AI narrows the decision space and the human makes the final judgment? → Tier 2: Collaborative.

Does the output inform human decisions without executing actions independently, with low-consequence, reversible outcomes? → Tier 1: Advisory.

If ambiguous, classify upward.

Step 2: Score Each Component (0-4)

For each of the five Human Layer components, select the level that best describes your current implementation. Use the evidence requirements to verify your assessment. The red flag indicators are designed to catch the most common self-scoring errors.

Component 1: Decision Gates

Level	Description	Evidence Required
0: Absent	No human checkpoint before consequential actions	N/A
1: Nominal	Policy requires approval; system does not enforce	Policy document exists; no system enforcement mechanism
2: Functional	System halts at gate; human must act to proceed	System logs showing gate activation; human input recorded
3: Structural	Gate logs action, timing, delta; bypass triggers exception	Telemetry dashboard; exception logs; delta reports
4: Adaptive	Monitors patterns; verification sampling; complexity-adjusted timing	Verification sampling results; complexity-adjusted baseline documentation; supervisor escalation logs

Red flags for over-scoring: If average gate approval time is under 5 seconds, you are likely at Level 2, not Level 3. If no exception has ever been triggered by a bypass attempt, the exception mechanism may not be functional. If you have never analyzed delta reports, you are not at Level 3. If you use fixed minimum

approval times rather than complexity-adjusted baselines, you are not at Level 4. If you do not deploy verification sampling (known-answer cases with intentionally incorrect AI recommendations), you are not at Level 4.

Component 2: Escalation Protocols

Level	Description	Evidence Required
0: Absent	System defaults to model output on all inputs	N/A
1: Nominal	Documentation describes escalation; system does not detect uncertainty	Escalation policy document; no system-level triggers
2: Functional	System detects uncertainty and routes to human	Escalation logs showing trigger events and routing
3: Structural	Logs track routing, response time, outcome; chains terminate in natural person; thresholds calibrated	Calibration records; chain-termination verification; response time analysis
4: Adaptive	System monitors escalation frequency; detects suppressed uncertainty; recalibrates	Recalibration records; suppression detection alerts; threshold adjustment logs

Red flags: If escalation has never triggered in production, the thresholds may be set too high or the detection mechanism may not be functional. If every escalation chain terminates in another AI system, you do not have escalation protocols under this specification.

Component 3: Accountability Structures

Level	Description	Evidence Required
0: Absent	No mapping between outputs and responsible humans	N/A
1: Nominal	Org chart names responsible roles; system does not enforce	Organizational documentation; no system-level identity linkage
2: Functional	Each consequential output maps to a named human	Output logs with human identity fields populated
3: Structural	Pre-assignment enforced; delegation explicit; contribution trace logs collaborator input; identity, authority, competency logged	Pre-assignment records; delegation chain documentation; competency verification logs; contribution trace records
4: Adaptive	System detects volume concentration; flags capacity exceedance; alerts on accountability gaps	Capacity monitoring dashboards; automated gap detection; personnel change alerts

Red flags: If the same individual is logged as accountable for more than 200 consequential decisions per day, the accountability may be nominal. If delegation has never been invoked, the delegation chain may not be functional. If you cannot identify, within 60 seconds, which human authorized a specific output from last month, you are not at Level 3. If committee-influenced decisions show no contribution trace, the primary anchor may be functioning as a moral crumple zone rather than an accountable decision-maker.

Component 4: Override Mechanisms

Level	Description	Evidence Required
0: Absent	No mechanism to reject or modify AI output at decision interface	N/A
1: Nominal	Override requires admin panel or technical intervention	Override capability exists; not accessible at decision point
2: Functional	Override accessible at decision interface	Interface screenshots; override action logs
3: Structural	Cognitive forcing functions present; element-level endorsement; justification logged; tested	Forcing function design documentation; test results; justification records
4: Adaptive	Monitors override frequency; detects bias/aversion; tests under degraded conditions; verification sampling confirms override engagement	Automation bias detection reports; stress-condition test results; frequency analysis; verification sampling results

Red flags: If the override rate is exactly 0% over any 30-day period, automation bias is likely present and Level 3 is not achieved. If the override has never been tested under time pressure or cognitive load, you are not at Level 4. If the override requires more than two clicks to execute, it may not be operationally accessible. If verification sampling cases (intentionally incorrect AI output) are approved without override, the override mechanism is architecturally present but functionally inactive.

Component 5: Trust Calibration Interfaces

Level	Description	Evidence Required
0: Absent	No reliability signals beyond model output	N/A
1: Nominal	Confidence score displayed; no external signals	Interface showing confidence score
2: Functional	Historical accuracy, boundary conditions, distribution shifts provided	Signal design documentation; examples of reliability indicators
3: Structural	RAIR/RSR calculated; disagreement rates logged; dependency monitored	RAIR/RSR calculation methodology; telemetry reports; dependency analysis
4: Adaptive	Calibration drift detected; RAIR/RSR divergence alerts; dual-track calibration for delayed ground truth; recommendation frequency adjusted	Drift detection logs; calibration adjustment records; proxy-to-outcome correlation data; agency preservation evidence

Red flags: If you display a confidence score and consider that sufficient, you are at Level 1, not Level 2. If you cannot produce RAIR and RSR numbers for the last 90 days, you are not at Level 3. If human performance has never been measured without AI assistance, you cannot assess dependency and are not at Level 4. If your domain has delayed ground truth (outcomes not known for weeks or months) and you have no proxy calibration methodology or retrospective true-up process, you are not at Level 4.

Step 3: Calculate and Assess

Sum the five component scores. This is your Human Layer Score (0-20).

Identify your minimum component score. This caps your assessment band.

Check against tier-specific minimums:

Tier 1. Minimum score: 10. No component below Level 2. Tier 2. Minimum score: 15. No component below Level 3. Tier 3. Minimum score: 18. No component below Level 3. At least three components at Level 4.

If your total score meets the threshold but any component falls below the tier minimum, the audit result is: Fail (Component Deficiency). Identify the deficient component and develop a remediation plan targeting the next maturity level.

If your total score falls below the tier threshold, the audit result is: Fail (Insufficient Maturity). The system does not meet the minimum standard for its risk tier.

If both conditions are met, the audit result is: Pass. Document the score, the date, and schedule the next assessment based on tier requirements (Tier 1: annually. Tier 2: semi-annually. Tier 3: quarterly, with continuous monitoring between assessments).

Step 4: Act on Results

For each component scoring below your tier requirement, the maturity criteria in Section 2 define what the next level requires. The gap between your current level and the required level is your engineering roadmap.

Priority order: address the lowest-scoring component first. The dependency graph means a weakness in one component degrades the effectiveness of all others. A Level 2 override mechanism undermines a Level 4 trust calibration interface, because the human may be well-calibrated in knowing when to override but structurally unable to do so effectively.

References

[1] Sheridan, T.B. and Verplank, W.L., "Human and Computer Control of Undersea Teleoperators," MIT Man-Machine Systems Laboratory, 1978.

[2] Parasuraman, R., Sheridan, T.B., and Wickens, C.D., "A Model for Types and Levels of Human Interaction with Automation," IEEE Transactions on Systems, Man, and Cybernetics, Part A, 30(3), 286-297, 2000.

[3] European Union, Regulation (EU) 2024/1689, "Artificial Intelligence Act," Article 14 (Human Oversight). Entered into force August 1, 2024. Fully applicable for high-risk systems August 2, 2027.

[4] Angelova, V., Dobbie, W., and Yang, C., "Algorithmic Recommendations and Human Discretion," NBER Working Paper No. 31747.

[5] Romeo, G. and Conti, D., "Exploring Automation Bias in Human-AI Collaboration: A Review and Implications for Explainable AI," AI & Society, Springer Nature, 2025. Systematic review of 35 studies (19,774 participants).

[6] Google DeepMind, retinal disease AI screening system deployed in NHS UK. Three-tier triage protocol.

[7] McLaughlin, B. and Spiess, J., "Complementary Algorithms," Stanford Graduate School of Business, 2025.

[8] Kahn, L., Probasco, E.S., and Kinoshita, R., "AI Safety and Automation Bias," Center for Security and Emerging Technology, Georgetown University, November 2024.

[9] International Organization for Standardization, ISO/IEC 42001:2023, "Information Technology: Artificial Intelligence: Management System."

[10] Frontiers in Big Data, "On the Purpose of Meaningful Human Control of AI," December 2022.

[11] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, January 2023.

[12] Infocomm Media Development Authority (IMDA), Singapore, "Model AI Governance Framework for Agentic AI," 2025.

[13] Kupfer, C. et al., "Automation Bias in Public Administration: An Interdisciplinary Perspective from Law and Psychology," Government Information Quarterly, 2024.

[14] Carnat, I., "Human, All Too Human: Accounting for Automation Bias in Generative Large Language Models," SSRN, March 2025.

[15] Schemmer, M. et al., appropriate reliance metrics (RAIR and RSR), as cited in research on AI-based decision support systems, Taylor & Francis, 2025.

[16] Wu, S., Liu, Y., Ruan, M., Chen, S., and Xie, X.Y., "Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation," Scientific Reports (Nature), 15(1), 15105, April 2025.

[17] Elish, M.C., "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction," Engaging Science, Technology, and Society, 5, 40-60, 2019.

[18] Daugherty, Paul R. and Wilson, H. James, "Human + Machine: Reimagining Work in the Age of AI," Harvard Business Review Press, 2018 (updated and expanded edition, 2024; ISBN 978-1647827205).

[19] Dell'Acqua, F., Ayoubi, C., Lifshitz, H., Sadun, R., Mollick, E., et al., "The Cybernetic Teammate: A Field Experiment on Generative AI Reshaping Teamwork and Expertise," Harvard Business School Working Paper No. 25-043, March 2025.

[20] Noureddine, A., "The Human Layer: Why the Most Critical Infrastructure in AI Isn't Artificial," Paper 1 in The Human Layer series, February 2026. DOI: 10.5281/zenodo.19119699.

[21] Noureddine, A., "The Human Layer Architecture: A Specification for Human-AI System Design," Paper 2 in The Human Layer series, March 2026. DOI: 10.5281/zenodo.19120077.

[22] Organisation for Economic Co-operation and Development, "OECD AI Principles," adopted May 2019, amended 2024. Adopted by 47 countries. Accountability principle requires traceability of datasets, processes, and decisions throughout the AI system lifecycle.

[23] Institute of Electrical and Electronics Engineers, IEEE 7000-2021, "Standard Model Process for Addressing Ethical Concerns During System Design," September 2021. Also published as ISO/IEC/IEEE 24748-7000:2022.

[24] Organisation for Economic Co-operation and Development, "G7 Hiroshima AI Process International Code of Conduct and Voluntary Reporting Framework," launched February 7, 2025.

[25] PwC, "Responsible AI and Internal Audit: What You Need to Know," 2026. Recommends adapting recognized frameworks (NIST AI RMF, ISO 42001, COSO) to address AI governance risks.

[26] ISACA, "Advanced in AI Audit (AAIA)" credential, 2025. First professional certification for AI audit competency, indicating market recognition that AI audit is becoming a distinct discipline.

Ahmad Noureddine is Co-Founder & CEO at Timer, the organizational memory layer. 25+ years building systems that put humans at the center of technology.

This is Paper 3 in The Human Layer series, published at ahmad.pt/research.

Paper 1: The Human Layer: Why the Most Critical Infrastructure in AI Isn't Artificial (DOI: 10.5281/zenodo.19119699) Paper 2: The Human Layer Architecture: A Specification for Human-AI System Design (DOI: 10.5281/zenodo.19120077) Paper 4: The Human Layer Economics: Capital Incentives and Organizational Automation Bias in AI