The Human Layer Architecture: A Specification for Human-AI System Design

Ahmad Noureddine

Timer · ORCID 0009-0003-1669-4553

2026-03-19 · DOI 10.5281/zenodo.19120076

The first paper in this series established the economic and institutional case for the Human Layer: the argument that AI systems designed to amplify human judgment outperform those designed for full replacement, supported by research across 1,500 organizations [18] and confirmed in post-GPT field experiments at Harvard Business School [19] and Stanford Graduate School of Business. [7]

That paper named five components of the Human Layer and promised a formal specification. This is that specification.

This specification applies to AI systems used in regulated, safety-critical, or trust-dependent environments, including finance, healthcare, public administration, and defense, but may also be applied to any commercial system where accountability for AI-assisted outcomes is required.

The five components are: decision gates, escalation protocols, accountability structures, override mechanisms, and trust calibration interfaces. Each is defined in this paper as a necessary architectural constraint. A system that omits any one of them cannot guarantee accountable human-AI interaction under regulatory or high-risk conditions. Each is independently required by at least one major regulatory framework now in force or approaching enforcement. The EU AI Act mandates human oversight for high-risk systems under Article 14. The NIST AI Risk Management Framework structures governance around four functions that map directly to these components. ISO/IEC 42001, the world's first AI management system standard, requires accountability assignment, risk treatment, and operational controls that align with all five.

None of these frameworks unifies the five components into a single coherent architecture. They define what organizations should do. They do not specify how to design the interface between human judgment and machine capability.

This paper provides that specification. It is not a policy recommendation. It is a design document for any builder, regulator, or institution operating AI in trust-dependent, regulated, or relationship-driven environments.

---

1. Decision Gates

A decision gate is a structurally mandated point in the system where human judgment is required before the system can proceed to the next action. It is not a notification. It is not a log entry. It is a hard dependency: the system cannot execute without human authorization.

The theoretical foundation for decision gates comes from the Levels of Automation taxonomy first proposed by Sheridan and Verplank at MIT in 1978. [1] Their framework describes ten levels ranging from "operator does it all" (Level 1) to "computer acts entirely autonomously" (Level 10). The critical range for trust-dependent systems is Levels 4 through 6: the system recommends an action and executes it only if the human approves (Level 5), or the system executes but the human retains veto power (Level 6). Parasuraman, Sheridan, and Wickens later refined this into four independent dimensions: information acquisition, information analysis, decision selection, and action implementation. [2] Each can be automated at different levels independently. Decision gates belong at the decision selection stage, the point where the system's analysis becomes an action with consequences.

The EU AI Act makes this concrete. Article 14(4)(d) requires that humans overseeing high-risk AI systems be enabled to "decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output." [3] For biometric identification systems, Article 14(5) goes further: any action based on the system's identification must be "separately verified and confirmed by at least two natural persons." [3] Two-person verification gates are not a recommendation. They are law.

The design principle is straightforward. Every consequential or irreversible action in the system must pass through a decision gate. The definition of "consequential" is domain-specific. In tokenized finance, it is the approval of an institutional-grade asset for market. In healthcare AI, it is a diagnostic recommendation that alters treatment. In criminal justice, it is a risk score that influences sentencing. The builder's job is to identify where in the system the consequences become irreversible and place the gate before that threshold.

This does not mean every system output requires human approval. A decision gate placed at every action point would destroy throughput and produce the same fatigue-driven rubber-stamping it is designed to prevent. Gates belong at irreversibility thresholds, the points where the cost of a wrong decision exceeds the cost of a delayed one. Routine, reversible, and low-variance outputs should flow without interruption. The architecture distinguishes between actions the system can safely execute autonomously and actions that require human authorization before the system proceeds. Getting that boundary right is the core design problem.

The failure mode when decision gates are absent is well documented. The system makes consequential decisions without human checkpoint. The canonical case is the Boeing 737 MAX MCAS system, which overrode pilot inputs based on a single sensor reading without providing pilots an accessible or intuitive mechanism to intervene. The sensor was the gate. It should have been a human.

But a study of pretrial bail decisions introduces a critical refinement. [4] When 90% of judges made override decisions no better than random, the decision gate was architecturally present but functionally decorative. A decision gate that produces rubber-stamp approvals is not a gate. It is a liability with a human signature attached. The architecture must be designed to produce high-quality human judgment at the gate, not merely the appearance of it. Decision gates must be enforced by the system, not by operator discipline. If the gate can be bypassed without triggering an exception, it is not a constraint. It is a suggestion.

---

2. Escalation Protocols

An escalation protocol is an explicit rule set that routes ambiguity-rich or edge-case scenarios to human judgment rather than allowing the system to default to model output. Where decision gates define the fixed points of human authority, escalation protocols define the dynamic triggers that activate human involvement in response to system uncertainty.

The critical design insight comes from the automation bias literature. A 2025 systematic review published in AI & Society examined 35 studies involving 19,774 participants and found that automation bias affects all users regardless of experience level. [5] High workload, time pressure, and task complexity increase overreliance on automated recommendations. Critically, explainability features alone did not mitigate the effect. Trust calibration feedback did not help either. [5] The implication is direct: if you leave escalation to the human's own judgment about when to question the system, the human will systematically fail to escalate precisely when escalation matters most.

The system must detect its own ambiguity and route accordingly. This is already operational.

Google DeepMind's AI system for eye disease diagnosis, deployed in the UK's National Health Service, uses a three-tier triage protocol. [6] The AI performs initial screening on retinal images and divides results into three categories: clearly normal, clearly abnormal, and uncertain. Cases classified as clearly normal require only rapid human review. Cases classified as uncertain escalate to specialist evaluation. The escalation trigger is embedded in the system architecture, not left to the discretion of the reviewing clinician.

A 2025 Stanford Graduate School of Business study on complementary algorithms applies the same principle to decision-making. [7] Rather than offering recommendations on every decision, the system offered selective recommendations only in cases where the human was likely uncertain or incorrect. People using this complementary approach made more accurate decisions than those using a purely predictive algorithm or no algorithm at all. The system decided when to intervene. That is the correct design pattern.

The Georgetown Center for Security and Emerging Technology published a 2024 analysis of automation bias in military systems that illustrates the catastrophic failure mode when escalation protocols are absent or overridden. [8] In the USS Vincennes incident, the AEGIS weapons system correctly identified a radar track, but the crew under extreme stress misclassified a civilian aircraft as a threat and did not escalate the ambiguous signal for additional verification. The system's information was accurate. The human escalation protocol failed under pressure. 290 people died.

The design principle: escalation protocols must be triggered by system-detected uncertainty thresholds, not by human self-assessment of their own confidence. The human decides what to do at the escalation point. The system decides when the escalation point is reached. And critically, every escalation chain must terminate in a decision gate controlled by a competent natural person, to use the EU AI Act's language. [3] An escalation from one AI model to a more capable AI model is not an escalation in the architectural sense. It is a lateral transfer within the machine layer. The escalation is only complete when it reaches a human with the authority and context to act on it.

---

3. Accountability Structures

An accountability structure is the architectural mapping between every consequential system output and the specific human or humans responsible for that output. It answers a single question: when this system produces a bad outcome, who is accountable?

This is not an organizational chart question. It is an architectural one. In complex AI systems, accountability diffuses naturally. The model was trained by one team, deployed by another, configured by a third, and operated by a fourth. When something goes wrong, the accountability distributes across "the algorithm," "the data," "the vendor," and "the operator" until no single human bears identifiable responsibility. ISO/IEC 42001 explicitly identifies this as the failure mode the standard targets: "diffused accountability, where no one owns AI risks until something goes wrong." [9]

The inverse failure is equally dangerous. Researcher Madeleine Clare Elish coined the term "moral crumple zone" to describe systems where a human is nominally positioned as the oversight layer but lacks the information, authority, or response time to meaningfully control the outcome. [17] When the system fails, the human absorbs the liability, functioning as a crumple zone that protects the institution while bearing the impact. Diffused accountability means no one is responsible. A moral crumple zone means the wrong person is responsible. The architecture must prevent both.

The distinction between accountability and responsibility matters here. A 2022 analysis published in Frontiers in Big Data articulates the difference. [10] Accountability requires that we know why the decision was A rather than B. It can sometimes be satisfied by technical means: an explanation, an audit log, a trace. Responsibility requires a human in the loop because, as the authors note, "we can't hold machines responsible in any meaningful sense." [10] These are different architectural requirements. A system can be accountable (traceable) without anyone being responsible (answerable). The Human Layer Architecture requires both.

The NIST AI Risk Management Framework structures this through its GOVERN function, which requires "organizational policies and practices" for AI oversight with "identified stakeholders responsible for security, compliance, and decision-making." [11] ISO 42001 Clause 5 requires top management to assign "accountability for AI-related decision-making" and integrate it into overall business strategy. [9] The EU AI Act Article 26(2) requires deployers to "assign human oversight to natural persons who have the necessary competence, training and authority." [3]

For agentic AI systems, where actions emerge dynamically from interactions rather than fixed logic, accountability structures face an additional challenge. Singapore's 2025 Model AI Governance Framework for Agentic AI addresses this directly: responsibility must be "clearly defined across multiple actors within and outside the organisation involved in the agent lifecycle." [12] The framework acknowledges that "human-in-the-loop remains effective over time notwithstanding automation bias" only when accountability is structural, not nominal.

The design principle: every consequential output in the system must have a pre-assigned human accountable for the decision and a pre-assigned human responsible for the outcome. These may be the same person. They may not be. But both must be identifiable before the output is produced, not determined after the outcome is known.

---

4. Override Mechanisms

An override mechanism is the operational capacity for a human to reject, modify, or halt AI system output without requiring technical intervention. The word "operational" is doing most of the work in that definition.

Many systems have theoretical override capabilities. An administrator can change model parameters. A developer can push a hotfix. A database administrator can modify a flagged record. None of these constitute an override mechanism in the architectural sense because none of them are accessible to the human operating the system at the point of decision. Override mechanisms must be available at the interface layer, not the infrastructure layer.

The EU AI Act is explicit. Article 14(4)(e) requires that humans be enabled to "intervene on the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure." [3] This is not metaphorical. The legislation envisions a literal mechanism, accessible to the oversight human, that halts or modifies the system's output in real time.

But placing a button on a screen does not solve the problem. The automation bias literature demonstrates that humans systematically defer to automated recommendations even when the option to override exists. A 2024 analysis of automation bias in public administration found that the EU AI Act's Article 14 requirements may be insufficient because they assume human oversight functions as intended. [13] A separate 2025 study proposed "cognitive forcing functions by design," meaning system-level mechanisms that require the human to actively engage with the decision rather than passively approve it. [14]

A simple example: instead of a single "approve" button, the override interface requires the human to select which specific elements of the AI's output they are endorsing and which they have independently verified. This is structurally different from clicking "OK." It forces engagement with the substance of the decision, not just the act of authorization.

The Georgetown CSET case study on the AEGIS weapons system illustrates the deeper design challenge. [8] Navy doctrine explicitly weights "decision-making towards the human" and maintains multiple qualified sailors in the identification loop "even when the system is in an autonomous mode." The override mechanism was present and doctrinally mandated. It still failed under extreme stress because the stress degraded the humans' capacity to exercise the override effectively. Override mechanisms must account for the reality that humans under pressure are not the same humans who designed the protocol.

The design principle: override mechanisms must be accessible at the decision interface (not buried in admin panels), must include cognitive forcing functions that prevent passive approval, and must be tested under degraded conditions, because the moment you most need an override is the moment human performance is

most compromised. An override that requires administrative intervention, specialized technical access, or navigation through multiple system layers does not qualify as an override mechanism under this specification.

---

5. Trust Calibration Interfaces

A trust calibration interface is the system mechanism that helps humans accurately assess when to trust AI output and when to question it. Of the five components, this is the most frequently misunderstood and the most frequently omitted from system design.

The common approach is a confidence score. The model outputs a prediction with 87% confidence, and the human uses that number to calibrate their trust. The research suggests this approach is inadequate.

The 2025 systematic review on automation bias found three distinct dimensions of trust that govern human-AI interaction: dispositional trust (the human's baseline tendency to trust automated systems), situational trust (trust shaped by the specific context and recent interactions), and learned trust (trust developed through repeated experiences with the system over time). [5] A confidence score addresses none of these dimensions. It describes the model's internal state, not the human's relationship to the model's reliability.

The EU AI Act recognizes this gap. Article 14(4)(a) requires that humans be enabled to "properly understand the relevant capacities and limitations" of the system. Article 14(4)(b) requires awareness of "the possible tendency of automatically relying or over-relying on the output." [3] The legislation demands that the system be designed to counteract the human's own cognitive biases, not merely to report the model's statistical properties.

Recent research on appropriate reliance proposes two metrics that better capture what trust calibration should measure. [15] Relative Positive AI Reliance (RAIR) indicates how often humans update their own incorrect decisions based on accurate AI advice. Relative Positive Self-Reliance (RSR) measures how often humans reject incorrect AI suggestions to rely on their own accurate judgments. A well-calibrated system maximizes both metrics simultaneously. This means the human both trusts the AI when the AI is right and trusts themselves when the AI is wrong. Neither blind trust nor blanket skepticism produces optimal outcomes.

A 2025 study in Nature's Scientific Reports adds another design constraint. [16] When workers transitioned from AI-assisted to solo work, intrinsic motivation decreased and boredom increased. A trust calibration interface that produces dependency, where the human cannot function effectively without the AI, has failed even if it maximizes short-term accuracy. The interface must preserve human agency and capability, not just human accuracy.

The Stanford complementary algorithm study offers the most actionable design pattern. [7] By providing recommendations selectively, only when the human was likely uncertain or incorrect, the system implicitly communicated when it had useful information and when it did not. The absence of a recommendation was

itself a calibration signal: the system was telling the human, "I don't have anything to add here. Trust your own judgment."

The design principle: trust calibration interfaces must address the human's relationship with the system (not just the model's confidence), preserve human agency and independent capability, and signal both when the AI is confident and when it is not.

Three formal requirements follow from this principle. First, trust calibration must provide reliability signals that are external to the model's internal confidence score. A model reporting its own certainty is self-assessment, not calibration. External signals include historical accuracy rates for similar inputs, known boundary conditions, and flagged distribution shifts. Second, the system must log human-AI disagreement rates and override patterns as part of trust calibration telemetry, enabling ongoing measurement of whether the human layer is genuinely engaged or passively deferring. Third, a system that maximizes decision accuracy while degrading the human's ability to perform independently is mis-calibrated by definition. Trust calibration that produces dependency has failed, regardless of its short-term performance metrics.

---

6. The Integrated Architecture

The five components described above are not a checklist. They are an interdependent system. Each component requires the others to function correctly.

Decision gates require escalation protocols to handle the cases that do not fit the gate's normal parameters. An asset verification gate works for standard submissions, but what happens when the submission falls outside the model's training distribution? Without an escalation protocol, the gate either rubber-stamps an edge case or blocks it without explanation. Neither outcome is acceptable.

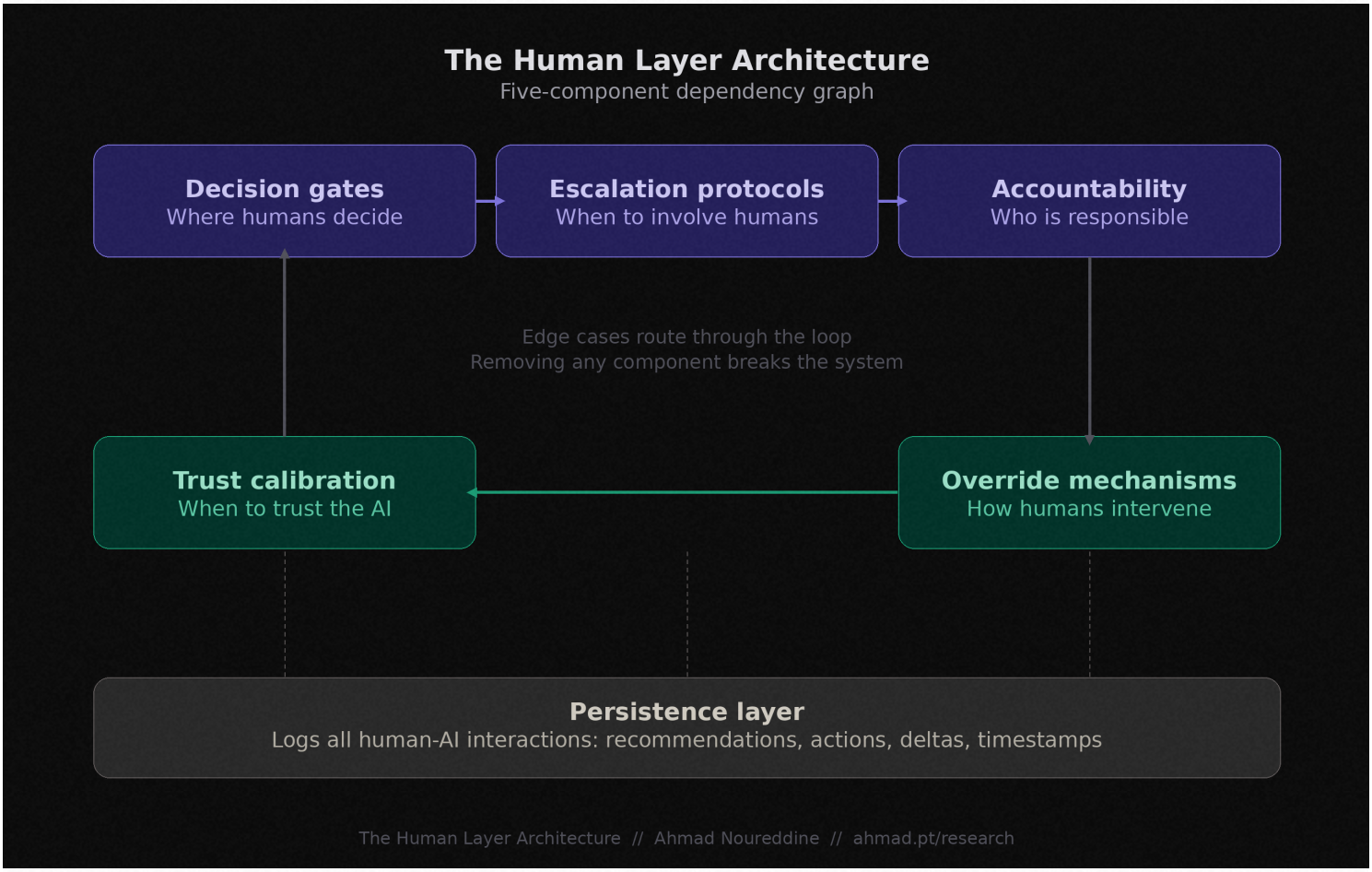
Escalation protocols require accountability structures to route the escalated decision to the right human. An uncertain diagnosis escalated from the AI triage layer needs to reach a specialist with the authority and competence to make the call. Without a pre-defined accountability structure, the escalation creates a queue, not a decision.

Accountability structures require override mechanisms to be enforceable. If a named human is accountable for a decision but cannot operationally modify the system's output, the accountability is nominal. They bear the responsibility without possessing the authority.

Override mechanisms require trust calibration interfaces to be exercised wisely. A human with the ability to override but no calibrated understanding of when to override will either never use the mechanism (automation bias) or use it indiscriminately (algorithm aversion). Both degrade system performance.

This dependency structure is why partial implementation fails. An organization that installs decision gates but ignores trust calibration will produce rubber-stamp approvals. An organization that builds override mechanisms but neglects accountability structures will produce overrides that no one is responsible for. The

architecture is only as strong as its weakest component. The five components form a dependency graph, not a menu. Removing any component breaks the guarantees provided by the others. The architecture must be validated as a whole. Component-level compliance is insufficient.



One structural requirement underlies all five components: a persistence layer that logs all human-AI interactions, including the system's recommendation, the human's action, the delta between them, and the time elapsed. Without this telemetry, the trust calibration metrics (RAIR and RSR) cannot be calculated, override patterns cannot be analyzed, and accountability cannot be traced after the fact. The logging layer is not a sixth component. It is the substrate on which the other five become measurable.

The convergence across regulatory frameworks reinforces this point. The EU AI Act Article 14 touches all five components in a single article. [3] The NIST AI RMF distributes them across its four functions: GOVERN (accountability), MAP (escalation), MEASURE (trust calibration), and MANAGE (decision gates, overrides). [11] ISO 42001 addresses them through Clauses 5 (accountability), 8 (operations, decision gates, overrides), and 9 (performance evaluation, trust calibration). [9] The regulatory landscape is converging on these five requirements. The only missing element is a unified specification that tells builders how to implement them as an integrated system.

This paper provides that specification. Paper 3 in this series will operationalize it into a scoring and assessment framework: The Human Layer Audit.

---

7. What This Architecture Makes Possible

The Human Layer Architecture is not a constraint on AI capability. It is the infrastructure that makes AI capability trustworthy enough to deploy at institutional scale.

A system built with all five components can operate in regulated markets without requiring external audit at every decision point, because the audit is embedded in the architecture. It can scale human judgment without scaling headcount, because the decision gates, escalation protocols, and trust calibration interfaces ensure that human attention is directed where it adds the most value. It can withstand regulatory scrutiny, because the accountability structures and override mechanisms create the traceability and control that regulators require.

A system built without these components can still function. It will be faster to build, cheaper to operate, and simpler to explain on a slide deck. But it will not earn institutional trust. It will not survive regulatory review. And it will not outperform, because the research cited throughout this series consistently shows that the collaboration architecture is the variable that separates augmentation from mere automation.

Systems intended for institutional deployment in regulated or trust-dependent environments must implement the Human Layer Architecture or provide an equivalent mechanism for decision gating, escalation, accountability, override, and trust calibration. The absence of these components is not a design choice. It is a structural deficiency that regulation is converging to prohibit and that market performance will penalize regardless.

References

[1] Sheridan, T.B. and Verplank, W.L., "Human and Computer Control of Undersea Teleoperators," MIT Man-Machine Systems Laboratory, 1978. Original 10-level taxonomy of automation.

[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. Extended Sheridan-Verplank into four independent automation dimensions.

[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. Found 90% of bail judges underperformed algorithm on overrides, but high-skill judges outperformed on both accuracy and racial fairness.

[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). Found automation bias affects all users regardless of experience; explainability alone does not mitigate it.

[6] Google DeepMind, retinal disease AI screening system deployed in NHS UK. Three-tier triage: clearly normal, clearly abnormal, uncertain. Dual review mechanism for uncertain cases.

[7] McLaughlin, B. and Spiess, J., "Complementary Algorithms," Stanford Graduate School of Business, 2025. Selective AI recommendations outperformed both full automation and unassisted human decisions.

[8] Kahn, L., Probasco, E.S., and Kinoshita, R., "AI Safety and Automation Bias," Center for Security and Emerging Technology, Georgetown University, November 2024. Case studies of AEGIS/USS Vincennes and Patriot missile automation failures.

[9] International Organization for Standardization, ISO/IEC 42001:2023, "Information Technology: Artificial Intelligence: Management System." First international AI management system standard.

[10] Frontiers in Big Data, "On the Purpose of Meaningful Human Control of AI," December 2022. Distinguishes accountability (knowing why decision was A rather than B) from responsibility (requiring a human who can be held answerable).

[11] National Institute of Standards and Technology, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, January 2023. Four core functions: GOVERN, MAP, MEASURE, MANAGE.

[12] Infocomm Media Development Authority (IMDA), Singapore, "Model AI Governance Framework for Agentic AI," 2025. Addresses accountability in dynamic agent systems where actions emerge from interactions rather than fixed logic.

[13] Kupfer, C. et al., "Automation Bias in Public Administration: An Interdisciplinary Perspective from Law and Psychology," Government Information Quarterly, 2024. Argues human oversight requirements may be insufficient due to automation bias.

[14] Carnat, I., "Human, All Too Human: Accounting for Automation Bias in Generative Large Language Models," SSRN, March 2025. Proposes cognitive forcing functions as design-level countermeasures against automation bias.

[15] Schemmer, M. et al., appropriate reliance metrics (RAIR and RSR), as cited in research on AI-based decision support systems, Taylor & Francis, 2025. Defines Relative Positive AI Reliance and Relative Positive Self-Reliance as dual metrics for trust calibration.

[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. Introduces the concept of human operators absorbing liability for autonomous system failures they could not meaningfully control.

[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). Based on research across 1,500 organizations. Found firms using AI for augmentation achieved 3x the performance improvement versus automation-only firms.

[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 (NBER Working Paper No. w33641). Pre-registered field experiment with 776 professionals at Procter & Gamble confirming augmentation effects in post-GPT context.

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 2 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 Paper 3: The Human Layer Audit: Measuring Accountability in AI Systems