VERACITY SOLUTIONS — EXECUTIVE RESEARCH

Human-in-the-loop won't save your business from AI mistakes

Why manual review fails at scale and how 50 years of supervisory control research points to a targeted control architecture

Author: Veracity Systems Architecture Group Published: 2026 Document ID: VER-WP-2026-01

Artificial intelligence is probabilistic, but consequential enterprise work demands strict accountability. To bridge this gap, organizations place qualified domain experts between model outputs and operational execution.

Over the past two years, enterprise leaders have converged on this human-in-the-loop model as an essential safeguard for AI deployment. The rationale is clear: consequential decisions should remain attributable to people who can exercise judgment, challenge the system, and be held accountable for the outcome. In some contexts, that principle is already embedded in law. Article 22 of the EU GDPR[1] restricts solely automated decisions with legal or similarly significant effects and, where such automation is permitted, requires safeguards including the right to human intervention. Regulators have made clear that this intervention must be meaningful. That requirement is no longer theoretical: in August 2026, the Dutch Data Protection Authority fined Uber €825 million or $966 million[2] over automated decisions affecting drivers' ability to work, including findings concerning inadequate human involvement. Uber is appealing the decision.

The Hidden Cost of Reconstruction

To understand why universal manual review fails at scale, consider what is required of a reviewer evaluating a 30-page AI-generated credit memorandum, insurance claim summary, or financial model.

The reviewer cannot simply read the text. They need access to the underlying data, calculations, scope, and assumptions to determine whether the output is reliable. In software testing, code can be evaluated against predefined test suites. Open-ended generative AI presents a less bounded review surface (see why model evaluation differs from output verification). Organizations must first define what constitutes an error and establish the standards against which outputs will be judged. NIST's AI Risk Management Framework[3] makes this distinction explicit, calling for organizations to identify appropriate metrics and evaluate systems against documented test sets and assurance criteria.

Some errors are straightforward. If a credit memorandum reports EBITDA of $42 million when the source financials show $38 million, the answer is wrong. Other evaluations require context and domain expertise. If the same memorandum describes 12% revenue growth as "strong," the reviewer needs a benchmark: historical

performance, peer performance, or an underwriting threshold. There is no automated test suite that can check this.

Once those standards exist, someone must perform the verification. This reconstruction burden creates a fundamental constraint on verification. Subtle errors can require the reviewer to rebuild much of the reasoning that produced the output in the first place. The burden compounds in multi-stage workflows, where an error introduced upstream can propagate through calculations, summaries, recommendations, and downstream decisions. This is because generation scales much faster than this reconstruction process. Across hundreds of portfolio companies or thousands of claims, AI can produce work products far faster than human reviewers can independently verify them.

> **The human reviewer becomes the operational bottleneck.**

As review queues grow, it is no surprise that review behavior changes, and the quality gate gradually becomes administrative: effective at catching surface-level defects, but less reliable at detecting structural, financial, or legal errors. As the bottleneck, human reviewers are pressured to move at a speed that no longer allows for rigorous analysis and review.

A 2025 Deloitte report for the Australian government[4] illustrates the limits of this safeguard. The 237-page report, prepared with generative AI assistance, contained nonexistent academic references and a fabricated quotation attributed to a Federal Court judgment. Deloitte later disclosed that the AI-generated material had been reviewed against source documents before publication, but while that process caught some errors, many other consequential ones slipped through. They were only identified after publication. The asymmetry is straightforward. Generating 237 pages is inexpensive. Verifying them is not.

Lessons from Fifty Years of Supervisory Control

The shortcomings of human review are not a novel byproduct of large language models. For over fifty years, cognitive psychologists have studied human supervision of automated systems in aviation cockpits, nuclear power plants, and industrial process control.

In 1978, while designing undersea teleoperators for the Navy, MIT and Stanford professors Thomas B. Sheridan and William L. Verplank introduced ten levels of automation, mapping how human roles shift from control to supervision as more decisions and actions are delegated to a machine. Five years later, cognitive psychologist Lisanne Bainbridge identified a fundamental paradox in this taxonomy in *Ironies of Automation*. As automation removes most of a human supervisor's routine involvement, it leaves them responsible for monitoring the system and intervening when something goes wrong. But reduced involvement makes it harder for supervisors to maintain the context needed to intervene effectively. Subsequent research by Sarter and Woods, Endsley and

Kiris, Skitka et al., and Parasuraman and Manzey documented related forms of automation bias, including cases in which operators defer to automated recommendations or fail to seek contradictory evidence.

This automation tension transfers directly to generative AI knowledge work. In a study of Boston Consulting Group consultants[5], researchers deliberately included a task outside GPT-4's capability frontier. On that task, consultants using AI were 19% less likely to reach the correct solution than those working without it. The model could still produce coherent, persuasive answers. They were still wrong. Human involvement therefore did not reliably catch the errors and prevent incorrect automated reasoning from reaching the final work product.

Architecting the Human Control Layer

Given the time and cost demands of human reviewers, companies and researchers have started looking to AI evaluators. Though research shows there is some merit to this approach in catching some errors, the ultimate solution to the review bottleneck cannot be to trust another probabilistic AI model to evaluate the first model's output. When verification depends on authoritative data, calculations, and explicit rules, humans must remain in the system.

To preserve human authority at scale, organizations must restructure the review task around a clear principle of minimizing the scope of human review. When conditions can be established mechanically, it should not consume scarce expert attention. Deterministic checks can verify numerical reconciliation, data feed freshness, schema constraints, and regulatory rule compliance before an AI draft reaches a human reviewer's hands at the release boundary (read how the release boundary operates in practice). Instead of requiring a human to define an error evaluation framework and hunt for those errors across an unbounded 30-page document, the control layer should narrow the review surface and present the reviewer with the evidence required to resolve the remaining uncertainty (see the case for independent verification at the authorization boundary).

Human judgment is the most valuable control resource an enterprise possesses. Systems must be designed to protect that judgment, not exhaust it.

Core Citations & Primary Literature

[^1]: **Sheridan, T. B., & Verplank, W. L. (1978).** *Human and Computer Control of Undersea Teleoperators*. MIT Man-Machine Systems Laboratory Technical Report. [^2]: **Bainbridge, L. (1983).** Ironies of Automation. *Automatica*, 19(6), 775–779. [^3]: **Sarter, N. B., & Woods, D. D. (1995).** How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors*, 37(1), 5–19. [^4]: **Endsley, M. R., & Kiris, E. O. (1995).** The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors*, 37(2), 381–394. [^5]: **Skitka, L. J., Mosier, K. L., & Burdick, M. (1999).** Does automation bias

decision-making? *International Journal of Human-Computer Studies*, 51(5), 991–1010. [^6]: **Parasuraman, R., & Manzey, D. H. (2010).** Complacency and bias in human use of automation: An attentional analysis. *Human Factors*, 52(3), 381–410. [^7]: **Dell'Acqua, F., McFowland, E., Mollick, E. R., Lakhani, K. R., et al. (2023).** *Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality*. Harvard Business School Working Paper 24-013. https://www.hbs.edu/faculty/Pages/item.aspx?num=64700[6]

References & Regulatory Citations

[1] **Article 22 of the EU GDPR** — https://gdpr-info.eu/art-22-gdpr/

[2] **fined Uber €825 million or $966 million** — https://techcrunch.com/2026/08/23/uber-faces-fine-of-nearly-1b-over-automated-driver-suspensions/

[3] **NIST's AI Risk Management Framework** — https://www.nist.gov/itl/ai-risk-management-framework

[4] **2025 Deloitte report for the Australian government** — https://www.businessinsider.com/deloitte-australia-issues-refund-ai-assurance-project-2025-10

[5] **study of Boston Consulting Group consultants** — https://www.bcg.com/publications/2023/how-people-create-and-destroy-value-with-gen-ai

[6] **https://www.hbs.edu/faculty/Pages/item.aspx?num=64700** — https://www.hbs.edu/faculty/Pages/item.aspx?num=64700