Designing the Human-AI Enterprise

You Cannot Scale Blind

Curtiss J. Shupe

08

You Cannot Scale Blind

Why human-AI execution requires a real telemetry layer across workflow, quality, risk, and business outcomes.

Most organizations will be able to tell you how many licenses they have purchased, how many copilots they have turned on, and how many prompts have been sent to their users. That's not performance measurement. That's adoption measurement. That will tell you that they have some sort of AI somewhere in their environment, but it will not tell you that their operating model is working. It will not tell you whether their human-AI execution is helping to make better decisions, where their override behavior is trending, whether their agents are trending beyond expected behavior, or whether their speed of execution is being bought at the expense of quietly degrading quality. The NIST AI risk management framework is pretty straightforward on this issue. Measurement is a core function of governance, not an optional after-action report. The Measure function should persist as risks, impacts, and knowledge change over time. That's all it takes to put to bed the idea that measurement of AI is really just a dashboard issue. It's not. It's an operating issue.

I use the phrase enterprise workforce performance telemetry to describe the measurement layer that ties together technical signals and workflow quality, human-AI interaction models, decision results, and executive awareness. The term is a design term in this series rather than an external term per se, but the need it addresses is well justified in terms of the source material provided. Microsoft's Work Trend Index: 2025 report portrays an emerging enterprise model in terms of intelligence on demand, human agent teams, and managers increasingly in charge of digital workers. Microsoft bases this on survey data from 31,000 workers in 31 countries, LinkedIn labor market trends, and Microsoft 365 productivity signals. Once human-AI teaming, supervised agents, improved decisions, and dynamic governance become normal operating conditions in an enterprise, it requires a way to see the whole system in operation in order to operate it effectively. Otherwise, executives are flying blind while believing they have visibility into the operation. This is not insight; it's a more attractive form of guesswork.

A good telemetry model has five levels. The first is activity telemetry. This is the lowest form of observability. It includes traces, metrics, logs, tool calls, task starts, retries, fallbacks, and workflow transitions. OpenTelemetry is good here because it has an official signal model, which helps to clarify what's going on in the technical world. Traces show how to execute through a system. Metrics are runtime measurements. Logs are just recording what's going on. This is important because, today, the way humans and AI systems interact is through software pathways, and these need to be specially instrumented. OpenTelemetry is the floor, not the house. It's good for the enterprise to be able to collect raw signals. It's just not enough.

The second layer is workflow telemetry. This is where the operation becomes visible: queue movements, approval times, exception paths, cycle times, handoffs, and escalation

Curtiss J. Shupe

frequencies. This layer will show if operations are really improving or just speeding up in uncontrolled ways.

The third layer is human-AI interaction telemetry. This is where many organizations still have weak visibility. How many times do users accept, edit, reject, or override the results of the AIs? Where do agents need human interaction? Where do users have too low of a trust level in the AIs, causing them to not use them at all, and where do they have too high of a trust level, causing them to rely on them too heavily? These patterns are important because they will show whether the operating model is really being executed as designed or if it's just being corrected in ways that management doesn't even see. The machine may be efficient on paper, but the humans are cleaning up the mess off the books.

The fourth layer is decision and quality telemetry. This is where the enterprise asks itself whether it's really getting improved decisions, error rates, consistency, rework amounts, or speed to resolution without sacrificing decision quality. While NIST doesn't recommend a specific scorecard for an enterprise, its approach to an AI system's lifecycle obviously includes ongoing assessment of system performance, impact, and risk changes. The fifth layer is business and governance telemetry. This is where we look at productivity, business resiliency, incidents, compliance signals, how the workforce is adapting, and business outcomes. International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 42001 provides another model that includes management system elements. ISO/IEC 42001 defines AI governance as including establishing, implementing, maintaining, and improving an AI management system. This means that there must be management action driven by telemetry. Telemetry that never changes a management action isn't governance – it's decoration. Perhaps it's pricey decoration. But it's still decoration.

This multi-layered approach is important because most organizations tend to think in terms of two buckets: infrastructure observability or executive dashboards. The former is operationally useful but not enough; the latter is rhetorically useful but not enough. The enterprise requires all the layers in between, and they require connection. This is the difference between knowing that a system is operating and knowing that it's operating correctly. This is also the difference between knowing that there's a weakness in a system and knowing that there's a weakness in a system's structure. If executives cannot connect tool usage to workflow behavior, human intervention, decision quality, and business outcome, they cannot govern the operating model they created. They can admire it, fund it, and discuss it at off-sites. They cannot govern it.

The workforce evidence makes the need sharper. PricewaterhouseCoopers' 2025 Global AI Jobs Barometer says organizations need a clear, data-based picture of skills gaps and reports that skills sought in AI-exposed jobs are changing faster than in other jobs. The World Economic Forum's Future of Jobs Report 2025 says skills gaps remain the biggest barrier to transformation for 63 percent of employers and that 39 percent of workers' key

skills are expected to change by 2030. Those aren't side notes. They mean the measurement challenge isn't only technical. Leaders have to see how work, judgment, escalation, and required capabilities are shifting in live operation. That's why this is a control-plane issue, not just a reporting issue. If you can't observe when a supervised agent drifts, you can't supervise it properly. If you can't see where humans repeatedly override a recommendation engine, you don't know whether the system is underperforming or whether trust is miscalibrated. If you can't connect speed gains to error patterns, escalations, and decision quality, you may be scaling fragility instead of performance. Organizations that can answer those questions have more than adoption metrics. They have the measurement layer required to manage human-AI execution at scale.

Curtiss J. Shupe