Closing the Collaboration Gap

A Behavioral Skill Framework for Human-AI Performance Improvement

Presented at the 2026 ISPI Performance Improvement Conference

Nashville, Tennessee • March 31, 2026

Sam Rogers

Founder & CEO, PAICE.work

Executive Summary

Organizations worldwide are investing billions to integrate AI into professional workflows, yet virtually none can answer a foundational question: how effectively do their people actually collaborate with AI? Training completion rates confirm attendance, not capability. Usage metrics confirm adoption, not quality. Policy compliance confirms awareness, not practice. The gap between what organizations measure and what actually determines AI outcomes represents one of the most significant unmeasured performance domains in the modern workplace.

This whitepaper introduces human-AI collaboration as a definable, measurable, and developable professional skill set and competency. It presents PAICE (People+AI Collaboration Effectiveness), a behavioral assessment framework built on five weighted dimensions (Performance, Accountability, Integrity, Collaboration, and Evolution) that measures what professionals actually do when AI systems produce errors, overconfident claims, and subtle failures. In HPT terms, these collaboration skills are treated as context-bound capacities to act. They are probabilistic levers that increase the likelihood of desired AI-supported accomplishments, but are neither prerequisites nor guarantees of success.

The framework draws on principles that the performance improvement field has championed for decades: that worthy accomplishments are the ultimate measure of performance, that observable behavior is the most reliable path to achieving them, that measurement must precede intervention, and that systemic analysis produces better outcomes than ad hoc training.

For performance improvement practitioners, this paper maps the PAICE Framework™ to established HPT models, including Gilbert's Behavior Engineering Model, the Mager and Pipe performance analysis flowchart, the Rummler-Brache Performance Framework, and ISPI's systematic HPT process, and traces its intellectual lineage through the evaluation science contributions of Will Thalheimer, Jack and Patti Phillips, and Robert O. Brinkerhoff. The paper demonstrates that AI collaboration measurement is not a departure from human performance technology but its natural next application, and concludes with a practical implementation approach for organizations seeking to baseline, develop, and govern this critical new skill domain and competency.

A New Performance Domain

Every generation of technology creates new performance demands. The industrial revolution required workers to operate machinery they had never seen. The information age required professionals to manage knowledge flows that moved faster than institutional memory. Each wave produced a period in which organizations invested heavily in new tools while underinvesting in the human capabilities needed to use them effectively.

Artificial intelligence is the current wave, and the pattern is repeating. Organizations are deploying AI tools across every professional function, including legal research, financial analysis, clinical documentation, insurance underwriting, software development, and customer communication, with the expectation that adoption will produce value. In many cases, it does. But adoption without measurement creates a performance domain that is invisible to management: the quality of the collaboration between human professionals and AI systems. As AI agents grow more capable and autonomous, both the human and the agent function as performers in the workflow, each contributing skills that affect the quality of the accomplished result. The human's skills remain primary and accountable, but the collaboration itself is the performance domain that requires measurement.

This is not a question of whether people *can* use AI. Most professionals have already demonstrated basic tool proficiency. The question is whether they use AI *well*: whether they maintain judgment when AI sounds authoritative, verify claims that sound plausible, and catch errors that are designed by their very nature to evade detection. AI does not fail loudly. It fails politely, with confident language and professional formatting that makes wrong answers indistinguishable from right ones.

> *The risk is not capability. The risk is calibration. Do your people know when to trust and when to verify? You are not measuring that yet.*

For performance improvement professionals, this should sound familiar. Thomas Gilbert recognized as early as the 1960s that the gap between exemplary and typical performance represents the single greatest opportunity for organizational improvement. In AI collaboration, that gap is not just unmeasured; it is unmeasurable by current methods. Training records, usage analytics, and policy attestations reveal nothing about what

happens in the critical moment when a professional encounters AI output that is subtly, confidently wrong.

The performance improvement field is uniquely equipped to address this challenge, because HPT has always understood that performance is behavioral, systemic, and measurable. The question is whether we will apply those principles to the defining skill domain of this decade.

Why Current Approaches Fail

The default organizational response to AI governance follows a well-worn playbook: write a policy, build a training program, track completion, and move on. Each of these steps is necessary. None is sufficient. The result is what practitioners might recognize as a classic HPT misdiagnosis: jumping to the Knowledge cell of Gilbert's Behavior Engineering Model without first examining whether the environment supports the desired behavior.

The Knowledge-Behavior Gap

Most AI governance programs measure declared intent: an employee's acknowledgment of a policy, their completion of an e-learning module, or their score on a knowledge assessment about AI risks. But in the fluid, high-speed reality of AI-assisted work, knowing the rules is a poor predictor of following them under pressure. A professional who can articulate the importance of AI verification on a quiz may still accept a confident-sounding hallucination without question when facing a deadline.

Gilbert codified this distinction in his 1978 book Human Competence, though the underlying ideas appeared in his earlier work throughout the 1960s and 1970s. He separated what people know from what people do, and further separated both from what people accomplish. HPT measures performance at the accomplishment level (the valued output), not at the knowledge level. A person who can recite verification principles but fails to detect actual AI errors has knowledge without behavioral skill. This is not a theoretical concern; it is the central pattern observed in behavioral AI collaboration assessments conducted across professional populations.

The Self-Assessment Problem

The gap is compounded by a measurement failure unique to AI. Unlike previous technologies, AI systems provide positive reinforcement regardless of user performance. Every response sounds helpful. Every output appears polished. The system never tells a user that they asked a poor question, accepted a fabricated statistic, or missed a critical error in the third paragraph. Over time, this creates an inflated self-perception of collaboration effectiveness. Professionals believe they are strong AI collaborators partly because AI has told them so, repeatedly and regardless of their actual behavior.

Self-reported AI readiness surveys inherit this bias. When asked whether they verify AI output, most professionals say yes. When their behavior is observed under controlled conditions, the gap between stated practice and actual practice becomes visible. This is not dishonesty; it is the predictable result of a system that provides no corrective feedback.

The Environmental Deficit

Gilbert's Behavior Engineering Model insists on a specific ordering: address environmental factors before individual factors. Yet most organizations attempting to improve AI collaboration skip directly to training (the Knowledge cell) without examining whether the workplace provides clear verification standards (Information), adequate verification tools and processes (Instrumentation), or incentives aligned with careful AI use (Motivation). When the environment does not support the desired behavior, training produces temporary awareness without lasting behavioral change.

Consider the professional who receives AI verification training on Monday and returns to a workflow on Tuesday that rewards speed and throughput with no mechanism for verifying AI output quality. The training addressed Knowledge. The workflow undermines Information, Instrumentation, *and* Motivation simultaneously. Gilbert would predict that the training will fail, not because the content was poor, but because the environmental supports are absent.

> *Policy tells people what to do. Capability measurement tells you whether they can actually do it. Most organizations have the first without the second.*

Human-AI Collaboration as a Definable Skill Set and Competency

If current approaches are insufficient, what would a rigorous skill set and competency model for human-AI collaboration look like? In HPT terms, these collaboration skills are context-bound capacities to act: probabilistic levers that increase the likelihood of desired AI-supported accomplishments, but are neither prerequisites nor guarantees of success. Not every failure to achieve an AI-supported outcome is a skill failure; environmental factors, tool limitations, and task complexity all mediate the relationship between skill and accomplishment. The PAICE Framework proposes that effective AI collaboration comprises five measurable dimensions, each weighted according to its contribution to safe, effective professional practice. The full framework specification, including twenty subscores, scoring methodology, and dimensional weighting, is detailed in the companion whitepaper, *PAICE.work: Making AI Collaboration Measurable, Teachable, and Governable (Rogers, 2025)*. What follows here is a summary oriented toward the performance improvement practitioner.

The PAICE Framework

The five dimensions (Performance, Accountability, Integrity, Collaboration, and Evolution) are not equally weighted. The weight structure reflects a deliberate design choice grounded in the evidence hierarchy described in the previous section. Dimensions that can be assessed through direct behavioral observation (injected-error tests) carry higher weights than dimensions assessed primarily through conversational inference. This means the score's center of gravity sits on the dimensions where behavioral ground truth is available.

Please note that **within the context of the PAICE Framework, "Performance" refers to tool-use proficiency**, not to performance in the HPT sense of valued accomplishments. The Performance dimension here measures prompt mastery, task framing, and operational efficiency, the skill of driving the tool well. In Gilbert's terms, this is a subset of the performer's behavioral repertory, not the accomplishment itself. The accomplished result, what HPT calls performance, is the outcome that all five PAICE dimensions contribute to collectively. The 10% weight reflects the fact that tool-driving skill, while necessary, is the most commoditized and fastest-changing component of the collaboration. Prompt techniques that are effective today may be obsolete in months as AI systems evolve. Verification judgment and domain expertise have longer half-lives.

Table 1: PAICE Framework: Five Dimensions of Human-AI Collaboration

Dimension	Weight	What It Measures
Performance*	10%	Prompt mastery, task framing, iterative refinement, operational value
Accountability	30%	Error detection, verification behavior, ownership of AI-assisted outputs, risk awareness
Integrity	25%	Factual grounding, logical consistency, contextual accuracy, domain-knowledge application
Collaboration	20%	Iteration quality, workflow design, shared control, specific feedback
Evolution	15%	Adaptive learning, critical evaluation, meta-awareness, experimental mindset

***** Performance = AI tool use subset of HPT performance

Why Accountability Carries the Highest Weight

Across assessment populations, Accountability is consistently the lowest-scoring dimension, typically 10–20 points below a professional's average across other dimensions. This is not surprising. AI systems present outputs with remarkable confidence and no hesitation. Human cognitive architecture is wired to trust authoritative-sounding information, especially under time pressure. Detecting subtle AI errors requires active skepticism, domain knowledge, and the cognitive effort to verify rather than accept.

A professional with strong Accountability skills and modest Performance is a better bet for safe outcomes than one with brilliant prompting skills and no verification habits. The weight structure encodes this principle: in a world where AI can generate plausible-sounding content instantly, the ability to verify, question, and maintain judgment is more valuable, and more rare, than the ability to prompt effectively.

This same logic explains why Evolution (15%) outranks Performance (10%) in the weight structure. AI capabilities are changing so rapidly that what a professional knows about effective AI collaboration today has a short shelf life. The professional who has mastered a specific set of prompting techniques but cannot adapt when the underlying model changes is less resilient than one whose current technique is modest but who actively evaluates new approaches, experiments with unfamiliar capabilities, and adjusts their collaboration patterns as the technology shifts. Evolution measures the rate of learning, not the current

state of knowledge. In a domain where the tools, failure modes, and best practices are moving targets, the capacity to learn and adapt is a more durable skill than any fixed technique. This is not a novel principle; it is the same logic that leads HPT practitioners to value adaptive expertise over routine expertise in any rapidly changing performance domain.

Behavioral Assessment Methodology

If human-AI collaboration is a behavioral skill set, it must be assessed behaviorally, not through knowledge tests, self-reports, or simulated scenarios with obvious correct answers. The PAICE Assessment™ is designed to observe actual collaboration behavior under realistic conditions.

Conversation-Based Assessment

Each assessment is a 25-minute conversation in which the participant brings a real task from their own professional context. They are not given a contrived scenario; they work on something that matters to them, using AI as they naturally would. This design choice serves two purposes. First, it eliminates the artificiality that makes most assessments poor predictors of on-the-job behavior. Second, it ensures that domain expertise can be applied. A lawyer working on a legal question can detect domain-specific errors that a generic scenario would never surface.

A natural question is whether a point-in-time assessment can substitute for continuous, on-the-job measurement of verification behavior. It cannot, and PAICE does not claim otherwise. The assessment establishes a behavioral baseline: it measures whether a professional can detect AI failures under realistic conditions when their domain expertise is engaged. Continuous measurement—random auditing of AI-assisted outputs, systematic verification checkpoints embedded in workflows—is the operational goal. But continuous monitoring cannot be designed effectively without diagnostic specificity. An organization must first know which failure modes its professionals miss, which dimensions need reinforcement, and where the gap between stated practice and actual behavior lives. PAICE baselines provide the diagnostic foundation that makes continuous monitoring meaningful rather than merely comprehensive.

Strategic Failure Injection

During the conversation, the system introduces controlled failures into the AI's responses: subtle errors that mirror the kinds of failures AI actually produces in professional settings. These are not obvious mistakes designed to be easily caught. They are confident-sounding claims with fabricated statistics, plausible but incorrect technical assertions, and contextual assumptions that the participant should recognize as flawed.

The participant's response to these injected failures provides the primary behavioral evidence for the assessment. Did they notice the error? Did they question it? Did they verify it independently? Or did they accept it and move on? This is the behavioral ground truth that knowledge tests cannot capture.

The Evidence Hierarchy

PAICE.work employs an explicit evidence hierarchy for scoring. Behavioral evidence (whether injected errors were caught or missed) is primary. Conversational evidence (what the participant says about verification practices, AI limitations, or collaboration principles) is secondary. When these conflict, behavioral evidence dominates.

This hierarchy addresses a specific pattern: professionals who discuss verification principles fluently while failing to apply them. A participant who articulates the importance of fact-checking AI output but accepts three fabricated statistics without question has demonstrated stated perception without behavioral skill. Their score reflects the misses, not the articulation.

> *High conversational fluency with missed tests reveals theoretical familiarity without behavioral skill. Low conversational fluency with caught tests reveals the skill that matters. PAICE.work scores the latter higher.*

Calibrated Skepticism, Not Paranoia

Productive AI collaboration requires calibrated trust: neither blind acceptance nor reflexive suspicion. A participant who challenges every AI response regardless of accuracy is not demonstrating Accountability; they are demonstrating a calibration failure. The assessment penalizes excessive false alarms (challenging correct output) because indiscriminate

skepticism is its own collaboration failure mode. In practice, it leads to verification fatigue, wasted effort, and the eventual abandonment of verification altogether.

The assessment measures whether a professional can distinguish situations that warrant verification from those that do not. This calibration, knowing when to trust and when to check, is the mature form of the skill set, and it is considerably harder than simply distrusting everything.

Calibration also includes a subtler skill: recognizing when AI output falls outside one's own domain expertise entirely. AI does not substitute for expertise; it builds upon it. A professional who cannot distinguish claims within their domain from claims beyond it will accept errors they lack the foundation to evaluate. This is not a separate competency from Integrity; it is what Integrity measures behaviorally. When participants accept domain-specific errors they lack the expertise to catch, the Integrity dimension surfaces the gap through observed behavior rather than self-assessment. The assessment does not ask professionals whether they know the boundaries of their knowledge. It reveals those boundaries through what they do.

Mapping to Established HPT Frameworks

Human-AI collaboration measurement is not a departure from human performance technology. It is the application of HPT's foundational principles to a performance domain that did not exist when those principles were articulated. The mapping is direct, and the alignment is substantive. These mappings are not retrospective justifications. The author's two decades of practice in learning and performance improvement, including work with organizations where these frameworks were operational, informed the design choices that produced PAICE.work. The frameworks cited here shaped the thinking that built the system; the mappings document that lineage explicitly.

Gilbert's Behavior Engineering Model

PAICE.work occupies a specific and critical role in Gilbert's framework: it provides the behavioral measurement that makes the performance gap visible and diagnosable. Specifically, PAICE focuses on performer-level skills of AI collaboration, evaluated through observable behavior, the individual-factors side of the BEM. This focus does not imply that every performance gap is a skill gap; it provides the diagnostic data needed to determine whether the gap is environmental or individual. Without valid measurement of current AI

collaboration performance, an organization cannot perform the cause analysis that the BEM requires. Is the gap an Information problem (unclear verification standards)? An Instrumentation problem (workflows that do not support verification)? A Motivation problem (incentives that reward speed over accuracy)? Or is it genuinely a Knowledge problem (professionals who do not know how to verify)?

Gilbert's ordering principle (address environmental factors before individual factors) remains as relevant to AI collaboration as to any other performance domain. But the ordering cannot begin without measurement. PAICE.work provides the diagnostic data that tells an organization which cells of the BEM to address and in what order.

Table 2: Mapping PAICE Dimensions to Gilbert's Behavior Engineering Model

BEM Cell	AI Collaboration Implication	PAICE Measurement
Information	Clear standards for when to verify AI output; feedback on AI error frequency	Cohort data reveals whether verification standards are understood and applied
Instrumentation	Workflows and tools that support verification; review checkpoints	Assessment identifies whether professionals have and use verification processes
Motivation	Incentives aligned with careful AI use rather than speed alone	Accountability dimension reveals whether professionals are motivated to verify
Knowledge	Training on AI error patterns, verification techniques, domain application	Score gap between conversational fluency and behavioral performance indicates knowledge-behavior disconnect
Capacity	Domain expertise sufficient to evaluate AI output quality	Integrity dimension reflects whether professionals apply domain knowledge to AI output
Motives	Personal commitment to maintaining professional standards with AI	Evolution dimension captures intrinsic motivation to improve collaboration practice

Mager and Pipe: Is This a Skill Problem?

Gilbert's BEM identifies which factors may be causing a performance gap. But the practitioner still needs a triage logic for routing the diagnosis to the right category of intervention. Robert Mager and Peter Pipe provided that logic in Analyzing Performance Problems (1970; revised editions 1984, 1997), one of the most widely used diagnostic tools in the performance improvement field. Their performance analysis flowchart guides the practitioner through a structured sequence of questions: Is the gap important enough to address? Is it a skill deficiency or an environmental deficiency? Could the performer do it if the stakes were high enough? Is there a simpler solution than training?

The central insight of Mager and Pipe is that training is the intervention of last resort, not the default. Their flowchart routes practitioners through environmental causes first (unclear expectations, inadequate feedback, misaligned incentives, missing tools) before arriving at the question of whether the performer lacks the skill to do the work. This ordering mirrors Gilbert's environmental-before-individual principle, but operationalizes it as a practical decision sequence rather than a diagnostic matrix.

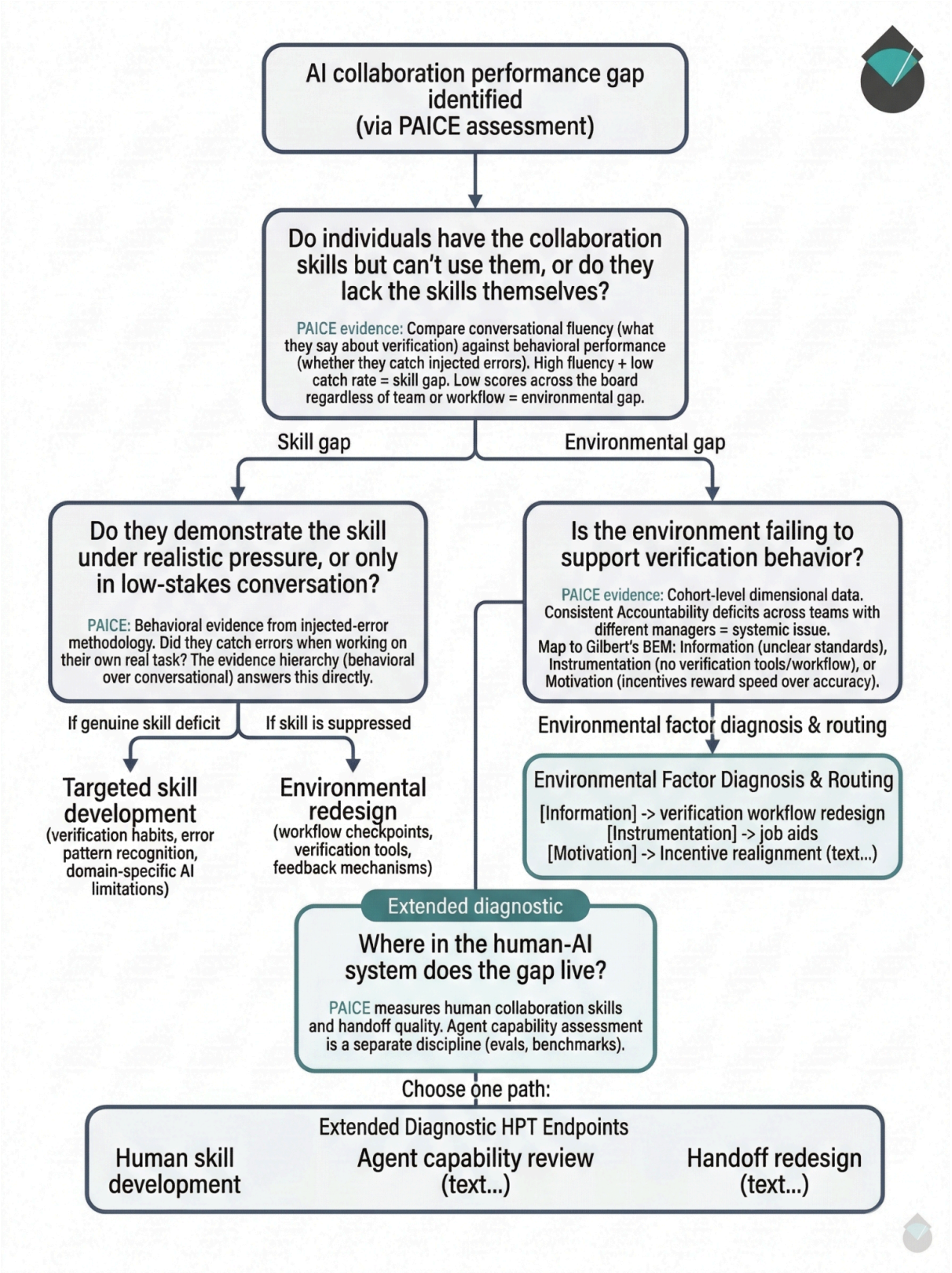
For AI collaboration, the Mager and Pipe questions have historically been unanswerable. Is the gap a skill deficiency? Without behavioral measurement, the practitioner has no evidence. Could they do it if the stakes were high enough? Without observing actual behavior under realistic conditions, no one knows. Is there a simpler solution than training? Without dimensional data showing whether the gap is in verification habits, domain application, or workflow design, every intervention is a guess. PAICE.work provides the behavioral evidence that makes each decision point in the Mager and Pipe sequence answerable with data rather than assumption. When dimensional scores show strong conversational fluency but weak error detection, the practitioner knows this is a skill gap, not an environmental gap. When cohort data shows consistent Accountability deficits across teams with different managers and different workflows, the practitioner knows training alone will not resolve it; the environment is not supporting verification behavior regardless of individual skill.

Mager and Pipe's canonical flowchart also introduces a question that takes on new significance in the context of People+AI collaboration: is the performance gap in the human performer, the agentic AI performer, or in the joint system they work together within?

PAICE.work currently measures the human side and the handoff quality, agent-side capability assessment is a separate discipline. But Mager and Pipe's triage logic provides the routing structure for distinguishing these causes, and PAICE data populates the decision points that make the routing evidence-based.

As such, this paper introduces a flowchart that builds upon the same logic for the Age of AI Agents. Though skills have traditionally been exclusively defined within the human domain, since the *AgentSkills* standard (https://agentskills.io/) was introduced in October 2025, this is no longer the case. The subject of how skills are defined and managed in agentic terms, and the ontological gaps and overlaps with existing human-first terminology, is something we are keenly interested to explore in a future paper.

Figure 1: PAICE-instrumented Performance Gap (next page)


AI collaboration performance gap identified (via PAICE assessment)
Do individuals have the collaboration skills but can't use them, or do they lack the skills themselves?
PAICE evidence: Compare conversational fluency (what they say about verification) against behavioral performance (whether they catch injected errors). High fluency + low catch rate = skill gap. Low scores across the board regardless of team or workflow = environmental gap.
Skill gap
Environmental gap
Do they demonstrate the skill under realistic pressure, or only in low-stakes conversation?
PAICE: Behavioral evidence from injected-error methodology. Did they catch errors when working on their own real task? The evidence hierarchy (behavioral over conversational) answers this directly.
Is the environment failing to support verification behavior?
PAICE evidence: Cohort-level dimensional data. Consistent Accountability deficits across teams with different managers = systemic issue.
Map to Gilbert's BEM: Information (unclear standards), Instrumentation (no verification tools/workflow), or Motivation (incentives reward speed over accuracy).
If genuine skill deficit
If skill is suppressed
Environmental factor diagnosis & routing
Targeted skill development
(verification habits, error pattern recognition, domain-specific AI limitations)
Environmental redesign
(workflow checkpoints, verification tools, feedback mechanisms)
Environmental Factor Diagnosis & Routing
[Information] -> verification workflow redesign
[Instrumentation] -> job aids
[Motivation] -> Incentive realignment (text...)
Extended diagnostic
Where in the human-AI system does the gap live?
PAICE measures human collaboration skills and handoff quality. Agent capability assessment is a separate discipline (evals, benchmarks).
Choose one path:
Extended Diagnostic HPT Endpoints
Human skill development
Agent capability review (text...)
Handoff redesign (text...)

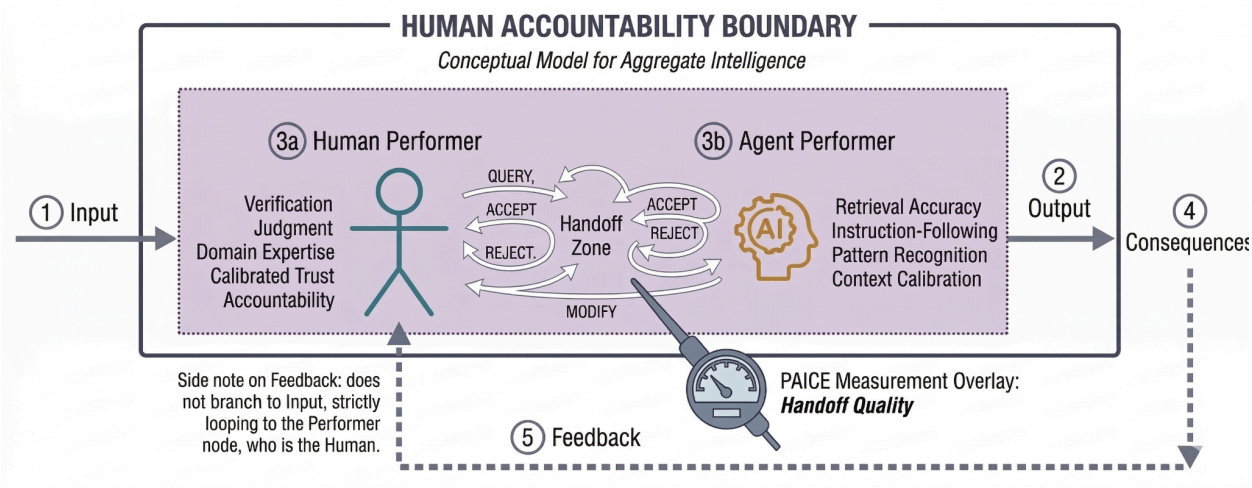
The Rummler-Brache Performance Framework

Rummler and Brache demonstrated that performance must be understood at three nested levels: Organization, Process, and Job/Performer. Most performance failures occur at process handoff points, the “white space” between functional silos that no single manager owns. Across all three levels, Rummler and Brache assess three needs: Goals, Design, and Management. PAICE operates primarily in the Job/Performer row, diagnosing whether collaboration goals are clear, whether the human-AI workflow is well-designed, and whether the collaboration is actively managed.

AI collaboration introduces a new category of white space: the handoff between AI output and human action. In this framing, both the human professional and the AI agent function as performers in the system, each with context-bound skills that contribute to or detract from the accomplished result. The agent’s skills (retrieval accuracy, instruction-following, calibration) are not the same as the human’s, but they are skills in the same sense: probabilistic capacities that affect outcomes without guaranteeing them. Human accountability remains primary, but treating the agent as an additional performer, rather than a passive tool, makes the handoff visible as a performance-critical interaction between two sets of capabilities. Every time a professional accepts, modifies, or rejects AI output, that handoff occurs. The quality of these handoffs is invisible to traditional management systems. PAICE.work makes it visible by measuring collaboration behavior at the Job/Performer level while generating cohort-level data that informs Process and Organization level interventions. Again, a full treatment of agent-side performance assessment within HPT frameworks is beyond this paper’s scope but represents a significant direction for future research at the intersection of human performance technology and AI systems design.

An organization that discovers its underwriting team has strong Collaboration scores but weak Accountability scores has identified a Process-level intervention target: the underwriting workflow needs verification checkpoints, not more collaboration training. This diagnostic precision, matching interventions to diagnosed causes, is exactly what the Rummler-Brache framework demands.

Figure 2: Human Performance System in AI Collaboration



ISPI's Systematic HPT Model

ISPI's five-phase HPT model begins with Performance Analysis: identify the desired state, measure the current state, and define the gap. For AI collaboration, this first phase has been largely skipped. Organizations have defined desired behaviors (through policies) without measuring actual behaviors (through assessment). The result is cause analysis conducted on an unmeasured gap, which is no cause analysis at all.

PAICE.work provides the measurement infrastructure for Phase 1. Once the gap is visible and dimensionally diagnosed, the remaining HPT phases (Cause Analysis, Intervention Selection, Implementation, and Evaluation) can proceed on an evidence base rather than assumptions. This is not a new methodology; it is the existing methodology applied to a new performance domain.

Table 3: PAICE Within ISPI's Systematic HPT Process

HPT Phase	Traditional Application	AI Collaboration Application
Performance Analysis	Define performance gap through observation and data	PAICE behavioral assessment baselines current AI collaboration capability
Cause Analysis	Use BEM or similar framework to identify root causes	Dimensional scores diagnose whether gaps are environmental or individual
Intervention Selection	Match interventions to diagnosed causes, not assumptions	Target training to knowledge gaps; redesign workflows for instrumentation gaps
Implementation	Deploy interventions with change management	Roll out targeted development with cohort-level tracking
Evaluation	Measure results against original gap	Reassess with PAICE to measure behavioral change, not just training completion

Intellectual Foundations in Evaluation Science

The PAICE Framework does not emerge from a vacuum. It stands on the shoulders of evaluation scientists whose work over the past four decades has progressively clarified what rigorous measurement of human performance requires, and why most organizations still fall short. Three contributions deserve specific acknowledgment, because each addresses a problem that PAICE inherits and extends into the AI collaboration domain.

Thalheimer: Measuring Decision-Making, Not Perceptions

Will Thalheimer's Learning-Transfer Evaluation Model (LTEM) provides an eight-tier hierarchy that distinguishes what most organizations measure (attendance, learner activity, and learner perceptions) from what actually matters: decision-making competence, task competence, and transfer to work. Thalheimer's core argument is that standard evaluation models actively mislead organizations by treating lower-tier metrics as evidence of learning effectiveness. Organizations that stop at learner satisfaction or knowledge recall believe they have evaluated capability when they have only measured comfort.

PAICE's evidence hierarchy reflects this principle directly. Conversational fluency about AI verification, analogous to Thalheimer's Tier 3 (Learner Perceptions), is not treated as

evidence of skill. A professional who sounds knowledgeable about AI collaboration but fails to catch injected errors is operating at the perception tier, not at the decision-making tier where competence actually lives. PAICE's injected-error methodology is designed to measure at Thalheimer's Tiers 5 and 6: can this person make correct decisions and perform effectively when AI output contains realistic failures? Because participants bring real professional tasks to the assessment rather than responding to contrived scenarios, the methodology incorporates elements of Tier 7 (Transfer), though a single assessment session cannot fully replicate the sustained on-the-job conditions that Tier 7 demands.

Phillips: Isolating Effects with Rigor

Jack and Patti Phillips' ROI Methodology, developed through the ROI Institute over three decades, introduced a discipline that most evaluation approaches lack: isolating the effects of an intervention from other influences. Their five-level framework extends Kirkpatrick's model by adding Return on Investment as a fifth level, but the deeper contribution is methodological: the insistence that organizations cannot claim impact without first controlling for confounding factors.

PAICE operationalizes this isolation principle within a single assessment session. By injecting known errors into AI responses and measuring whether the participant catches them, the assessment creates a controlled evaluation condition inside a naturalistic interaction. The test-catch rate is the isolation mechanism: it separates verified behavioral skill from domain knowledge, conversational style, and self-perception. When a professional catches three of four injected errors, that is not a self-report or a manager's impression. It is observed, isolated behavioral evidence, the kind of evidence the Phillips methodology demands.

Brinkerhoff: Studying What People Actually Do

Robert O. Brinkerhoff's Success Case Method (SCM), first articulated in 1983 and fully developed in his 2003 work, challenged the field to stop averaging across populations and start studying the extremes. His insight was that a program producing a few spectacular successes and many failures is a fundamentally different situation from one producing uniform mediocrity, but average-based evaluation methods show them as identical. The actionable information lives at the extremes: what did the most successful participants do differently, and what prevented the least successful from benefiting?

PAICE's scoring architecture reflects this principle. The tier system, from Constrained through Exceptional, is designed to differentiate meaningfully between distinct performance profiles, not to produce a smooth normal distribution. Roles that are not actively engaged in the design of new AI systems will likely never benefit from higher PAICE scores. The Exceptional tier is deliberately difficult to reach, preserving headroom as population capability improves over time. And Brinkerhoff's emphasis on identifying environmental enablers and barriers maps directly to PAICE's organizational application: cohort-level data reveals not just who is struggling but what conditions in the work environment are enabling or preventing effective AI collaboration.

> *Gilbert taught us that performance is about worthy accomplishments, and that observable behavior is how we get there. Mager and Pipe taught us to diagnose before we prescribe. Thalheimer taught us to measure decisions, not perceptions. Phillips taught us to isolate effects. Brinkerhoff taught us to study what people actually do. PAICE.work applies all five principles to the defining skill and competency challenge of this decade.*

Privacy-Preserving Measurement

Behavioral assessment of professionals in regulated industries creates an inherent tension. Organizations need cohort-level data to demonstrate risk mitigation to regulators. Individuals need development insight without their scores becoming a liability, weaponized by employers for performance management, hiring decisions, or workforce reduction.

PAICE.work resolves this tension through privacy by architecture, not by policy. The system is designed so that individual scores cannot be disclosed to or reverse-engineered by enterprise buyers. This is not a contractual promise; it is a structural constraint.

Architectural Privacy Guarantees

Individual assessment records are not retained in linkable form after delivery to the individual. Enterprise buyers receive only cohort-level aggregations (distributions, percentiles, trend lines) with no individual mapping. Conversation content is processed in real time during assessment and is never stored in production environments. Only final scores persist, and those scores are not linked to identifiable individuals.

This architecture means that even under legal or commercial pressure, individual-level results cannot be extracted from the system, because they are not in the system. The

privacy guarantee can be mathematical, not just contractual. The full technical architecture, including TEE-protected inference, on-chain score attestation, and regulatory framework mapping, is detailed in the companion whitepaper, *Verifiable Human-AI Collaboration: Privacy-Preserving Assessment with Cryptographic Integrity (Rogers, 2026)*.

Why This Matters for Regulated Professionals

The target users for PAICE.work are professionals in regulated industries: lawyers, clinicians, financial advisors, insurance underwriters, cybersecurity professionals. These individuals are personally licensed and personally liable, not just their employers. A lawyer who over-relies on AI output without verification risks their license. A clinician who accepts AI recommendations uncritically risks patient harm. These professionals need to develop their AI collaboration skills, but they also need assurance that an assessment will not be used against them.

For the organizations employing these professionals, privacy-preserving measurement resolves a governance dilemma. They can demonstrate to regulators that they are proactively assessing and developing AI collaboration capability across their workforce without creating an individual-level liability record. The cohort data answers the regulatory question (“Are our people using AI safely?”) without answering the employer question (“Which specific person scored lowest?”).

From Measurement to Development

Assessment without development is diagnosis without treatment. PAICE is designed not only to measure AI collaboration skills but to provide the diagnostic specificity needed for targeted performance improvement.

Individual Development

Each participant receives dimensional feedback showing their profile across all five PAICE dimensions. Because the assessment is behavioral rather than knowledge-based, this feedback reveals patterns the individual may not be aware of: a tendency to accept AI claims without verification, difficulty providing specific feedback to improve AI output, or a pattern of deferring to AI on matters within their own expertise. These patterns are actionable. A professional who learns they have strong Collaboration but weak

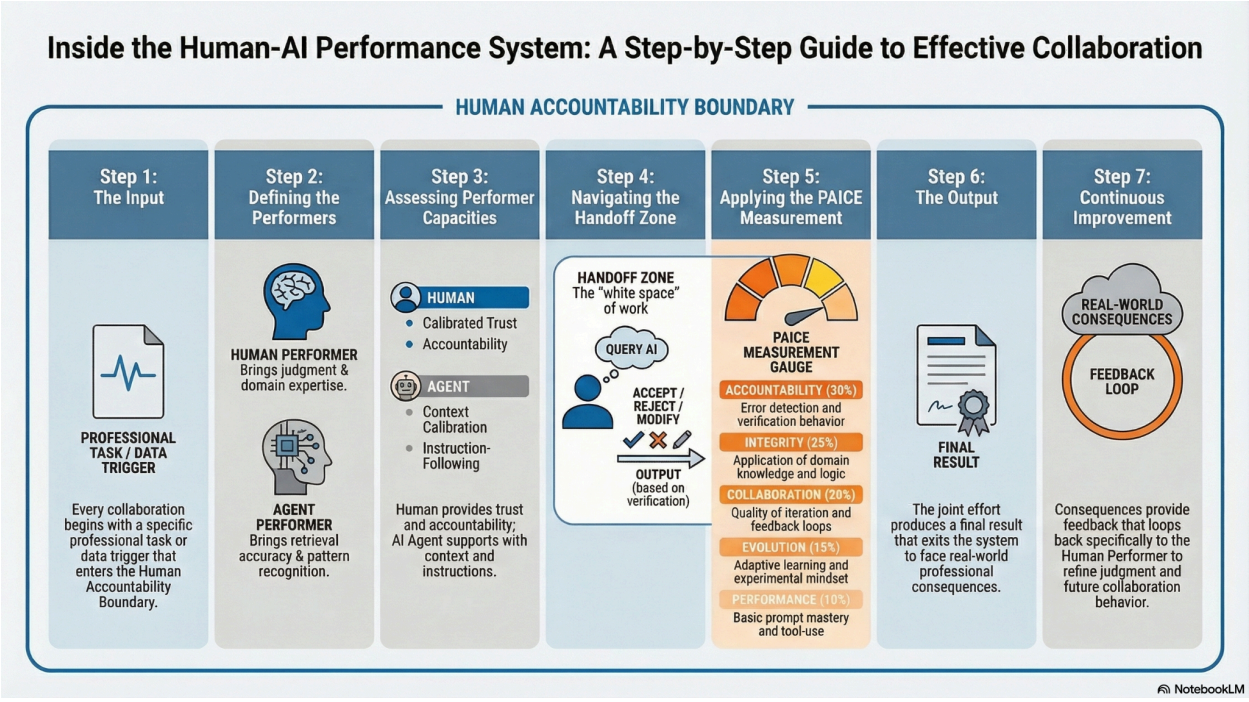
Accountability knows exactly where to focus skill development effort: building systematic verification habits.

Organizational Intervention Design

Cohort-level data enables evidence-based intervention design. When an organization can see that its risk assessment team has strong Performance and Evolution scores but weak Accountability and Integrity scores, the intervention is clear: the team knows how to use AI tools and is willing to adapt, but they are not catching errors or applying domain expertise to AI output. The intervention is not more AI training; it is verification workflow redesign, supported by job aids and perhaps domain-specific error pattern education.

This diagnostic precision prevents the most common intervention failure: treating every AI collaboration skill gap with the same generic awareness training. Generic training addresses the Knowledge cell of the BEM regardless of whether Knowledge is the actual cause. PAICE.work data enables the HPT practitioner to match the intervention to the diagnosed cause, which is the fundamental principle of the field.

Figure 3: A Human-AI Performance System Approach



Implementation Approach

A practical implementation follows a structured sequence that mirrors the HPT process:

Weeks 1–2: Baseline Assessment. Cohort of 20–100 professionals completes their PAICE assessment. No system integration required. No personal data collected. Participants bring their own professional context to a 25-minute conversation-based assessment. Individual results and axis of improvement are delivered immediately to participants; cohort data is aggregated for organizational analysis.

Week 3: Diagnostic Analysis. Organizational analysis identifies patterns across the five dimensions: which teams show verification gaps, where Accountability diverges from other dimensions, which cohorts demonstrate the strongest and weakest AI collaboration profiles and which display the most calibrated confidence. This analysis maps to BEM cells and identifies targeted intervention opportunities.

Week 4: Executive Readout and Intervention Planning. Findings are presented in a governance-ready executive report with specific recommendations: which teams are safe to scale AI adoption, which need targeted support, and what interventions will produce the highest-impact improvement based on the diagnosed causes.

Ongoing: Reassessment and Evaluation. Periodic reassessment measures behavioral change against the original baseline, completing the HPT evaluation cycle. Organizations can demonstrate not just that they trained people, but that training produced measurable behavioral improvement in AI collaboration.

Performance Improvement's Moment to Shine

Human-AI collaboration is not a technology problem with a technology answer. It is a performance problem (behavioral, systemic, and measurable), a skill challenge that demands the methodological rigor that performance improvement professionals have developed and refined for over sixty years.

The principles that this community has championed are precisely the principles that AI governance needs now. Worthy accomplishments are the measure of performance, and observable behavior is the most reliable path to achieving them. Measurement must precede intervention. Environmental analysis must precede individual development. Evidence-based cause analysis must replace assumption-driven training. These are not new

ideas. They are established, proven ideas that have not yet been applied to the defining performance challenge of this decade.

The PAICE Framework demonstrates that human-AI collaboration can be defined, measured, and developed with the same rigor that HPT brings to any other performance domain. It provides the behavioral measurement infrastructure that enables the full HPT cycle, from performance analysis through cause analysis, intervention selection, implementation, and evaluation, grounded in observed behavior rather than declared intent.

The organizations that will navigate AI adoption successfully are not those with the most sophisticated AI tools. They are those that can measure and develop the human skills needed to use those tools safely, effectively, and with maintained professional judgment. That capability is a skill set and a competency. It can be assessed. It can be improved. And the performance improvement field is uniquely positioned to lead its development.

Learn More About PAICE.work

Take the assessment: PAICE.work (free for individuals)
Organizational programs: PAICE.work/baseline
Contact: sales@PAICE.work

PAICE.work is a Public Benefit Corporation dedicated to making AI adoption measurable, teachable, and governable.

Disclaimer

This whitepaper is provided for informational purposes only and does not constitute professional, legal, or regulatory advice. The PAICE framework and assessment methodology described herein are based on research and development current as of March 2026. Organizations should consult their own compliance, legal, and risk management professionals before implementing any assessment program. The views expressed are those of the author and do not represent the positions of the International Society for Performance Improvement.