# PAICE.work
## Making AI Collaboration Measurable, Teachable, and Governable

*A framework for responsible AI adoption across individuals, teams, and organizations.*

Vision & Partnership Whitepaper

**Research Preview 2025.11**
**Author:** Sam Rogers, PAICE.work PBC
**Date:** November 12, 2025
**Version**: 3.1

**Audience:**
Learning, People, AI, Risk, and Operations leaders evaluating AI collaboration measurement

## 1. Executive Summary

Organizations are accelerating AI adoption without a reliable way to measure the single variable that determines success or failure: **how effectively people collaborate with AI systems.**

Traditional metrics track activity (usage rates, training completions, productivity outputs, etc.) but they do not measure competence. The result is predictable: unmeasured risk, wasted enablement budgets, compliance exposure, and blanket policies that limit everyone to protect against the few.

**PAICE.work** (People + Artificial Intelligence Collaboration Effectiveness, pronounced "pace-dot-work") provides a new layer of measurement infrastructure. It's a behavioral scoring framework that quantifies *collaboration quality* rather than activity. Through a 20-minute adaptive assessment, PAICE.work observes what people do when AI fails, misleads, or biases outputs. Those responses reveal true capability: verification, correction, ethical reasoning, and adaptation.

The output is a **0–100 PAICE Index score** and detailed guidance mapped to the five core dimensions:

1. **Performance:** Clarity and efficiency under variable conditions
2. **Accountability:** Verification and traceability of actions
3. **Integrity:** Consistency and factual grounding
4. **Collaboration:** Workflow design and iteration quality
5. **Evolution:** Reflection, adaptation, and continuous improvement

Each dimension aligns with international standards such as **NIST AI RMF** and **ISO/IEC 42001**, giving organizations a defensible way to benchmark readiness and improvement.

## 1.1 Current Status

PAICE.work is operational today for **individual assessments** available at https://paice.work. The scoring engine, conversational evaluation methodology, and framework logic are fully functional and already producing valuable insights.

This release is designated **Research Preview 2025.11** to emphasize transparency about validation: the system works, but formal benchmarking and longitudinal studies are now commencing.

## 1.2 Call for Pilot Partnerships

PAICE.work is seeking **3-5 enterprise or academic partners** to participate in structured pilots during Q1–Q2 2026. Pilot partners will:

- Assess baseline employee or student cohorts (10–100 participants)
- Conduct pre-/post-training comparisons for AI-readiness programs
- Share anonymized behavioral data for cross-industry benchmarking
- Co-author case studies linking measured capability to real-world outcomes

Partners receive:

- Free access to all assessments for pilot participants
- Early access to team analytics features (Q1 2026)
- Co-marketing and thought-leadership visibility
- Direct influence on the next-generation PAICE roadmap

**Objective:** establish the empirical foundation for measurable, auditable AI collaboration.

**Position:** PAICE.work is not selling a finished product, it's building shared infrastructure for the measurement era ahead. If you need to assess which AI tools your people are ready for, how to roll them out, and how to mitigate the associated risks, PAICE.work is the answer.

**Key Insights**

- Simple FICO-like score, deep assessment methodology
- Individual assessments live, free, and fully-functional today
- Cohort (Team, Training) & Enterprise pilots commencing soon

---

## 2. The Governance Gap: Why AI Adoption Outpaces Oversight

Organizations have spent decades building controls for financial, operational, and cybersecurity risk. Yet as AI proliferates, the most common source of failure isn't the technology, it's human behavior under uncertainty.

AI capabilities are evolving faster than most governance and learning systems.
Without shared metrics, "responsible AI" remains a statement of intent rather than evidence of performance.

### 2.1 The Measurement Blind Spot

Current reporting focuses on **activity, not capability**:

- Prompts or queries per day
- AI-feature activation rates
- Training module completions
- Chatbot sessions initiated

None of these reveal whether users recognize hallucinations, detect bias, or know when to override the machine. Failures surface only after damage is done:

- A legal brief citing fabricated cases
- A hiring model reinforcing bias
- A financial report issued without verification
- A chatbot leaking confidential data

In each example, metrics looked healthy. The failure was *behavioral.*

## 2.2 Why Traditional Metrics Miss the Risk

| Approach | What It Measures | What It Misses | Resulting Risk |
|---|---|---|---|
| **Knowledge tests** | Recall of AI facts | Judgment under pressure | False confidence |
| **Productivity metrics** | Output volume | Quality and accuracy | Liability exposure |
| **Self-assessments** | Perceived skill | Actual competence | Dunning–Kruger blind spots |
| **Training completions** | Content exposure | Behavioral change | Wasted investment |
| **Usage analytics** | Adoption rates | Collaboration quality | Undetected risk accumulation |

## 2.3 The Organizational Consequences

1. **Unmeasured Risk** AI-related failures are discovered post-incident. Risk teams cannot quantify exposure, and compliance leaders cannot prove competence to regulators or boards.
2. **Wasted Enablement Spend** L&D cannot demonstrate ROI, target interventions, or benchmark improvement without behavioral data. Training success is reduced to completion percentages.
3. **One-Size-Fits-None Policies** Policies designed for the least-prepared users constrain innovation for the most capable ones. Graduated access models—essential for safe scaling—are impossible without a competence baseline.
4. **Emerging Compliance Gaps** Frameworks like the EU AI Act, NIST AI RMF, and ISO/IEC 42001 emphasize *competence verification*. “We completed the training” no longer satisfies auditors asking, “Can your people use AI safely?”

## 2.4 The Path Forward

Organizations need a system that:

- Observes *behavior* under AI-failure conditions
- Produces quantified capability metrics for risk and compliance reporting
- Identifies dimension-specific gaps for targeted learning investment
- Enables graduated access based on demonstrated proficiency
- Provides before-and-after data to validate training impact

**PAICE was built for exactly this purpose**, to close the measurement gap between AI activity and AI accountability, giving leaders a defensible way to prove readiness *before* incidents occur.

**Key Insights**

- Adoption is fast, oversight is slow.
- No shared measure exists for AI collaboration quality.
- Measurable governance enables accountability and innovation.

# 3. What We Measure: The PAICE Framework

Organizations don't fail with AI because they lack knowledge. They fail because people respond inconsistently when AI produces *confidently wrong* answers.

PAICE measures those responses.

## 3.1 The Five Dimensions of Collaboration Capability

PAICE quantifies how individuals behave under AI failure conditions across five interdependent dimensions.

| **Dimension** | **Measures** | **Why It Matters** | **Business Risk if Low** |
|---|---|---|---|
| **Performance** | Task definition, prompt clarity, efficiency | Translates goals into usable AI output | Productivity loss |
| **Accountability** | Error detection, verification, bias awareness | Prevents AI-related harm | Compliance & legal exposure |
| **Integrity** | Factual grounding, logic, transparency | Keeps outputs trustworthy | Reputational damage |
| **Collaboration** | Iteration quality, workflow alignment | Improves human-AI synergy | Process drag |
| **Evolution** | Reflection, pattern learning, adaptation | Builds organizational learning loops | Stagnation |

Each dimension maps to controls in **NIST AI RMF** and **ISO/IEC 42001**, giving compliance teams a vocabulary already recognized by regulators. These tier boundaries are based on Research Preview 2025.11 data and will be refined through pilot partnerships (see *Section 6*).

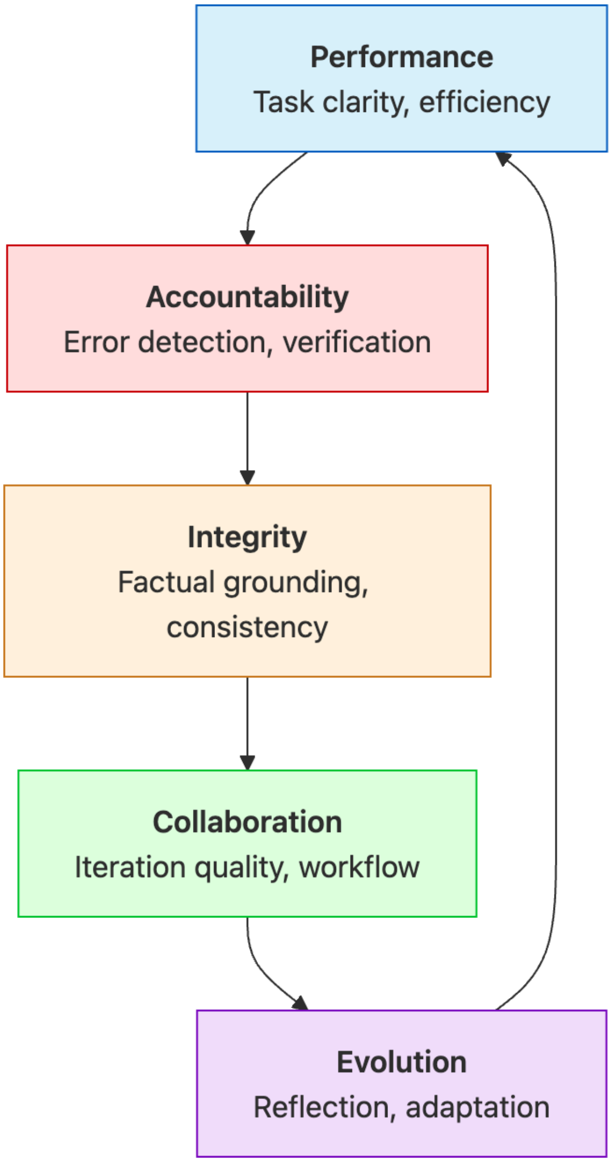


Figure 3.1 *The Five PAICE Dimensions*

## 3.2 The Accountability Gap

Preliminary assessment data show **Accountability** trailing other dimensions by 10-20 points, a signal that verification under pressure is the hardest skill to master.

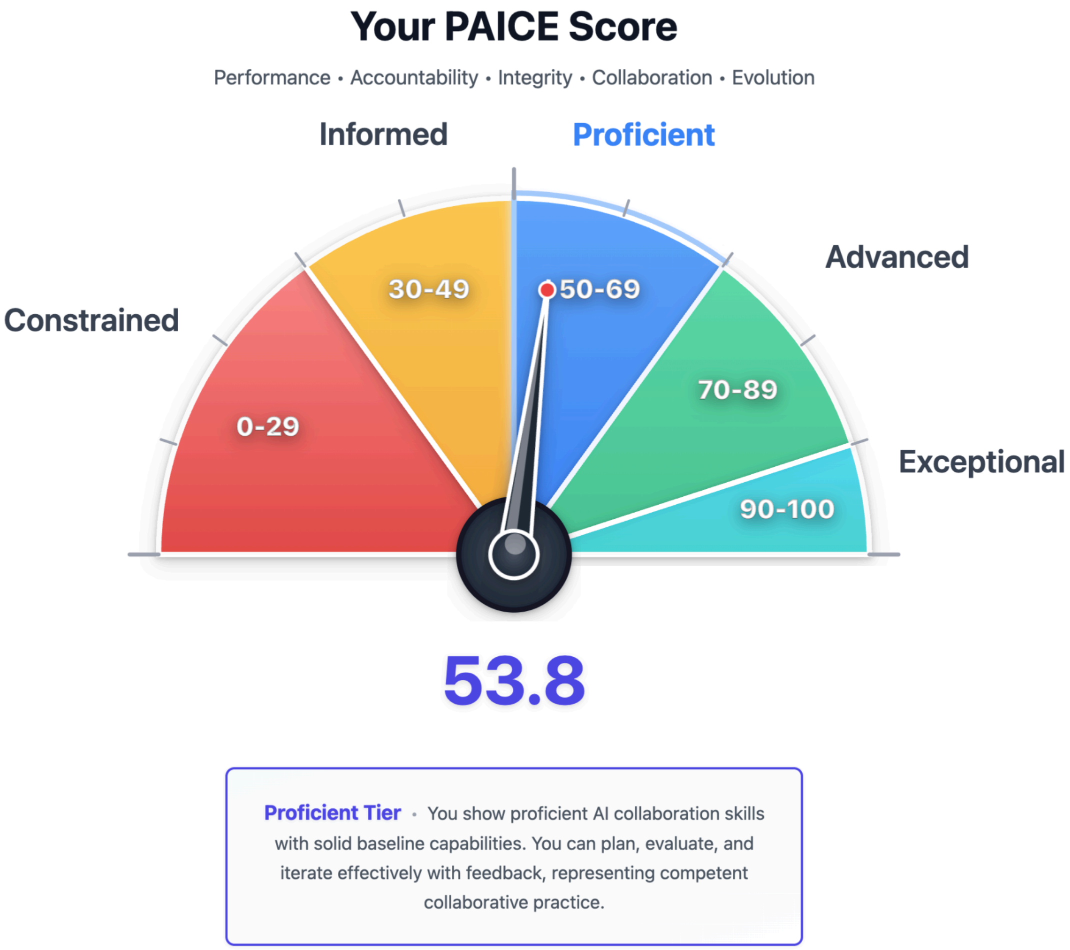
AI delivers outputs with authority but no uncertainty cues. Catching its errors requires sustained skepticism and time most users don't budget.

That gap is where organizational risk concentrates. A single unverified AI output can trigger legal, reputational, and financial damage.

Accordingly, **Accountability carries 30 percent of total PAICE weighting**, the largest share among dimensions.

## 3.3 Tier System for Interpreting Scores

The PAICE Index (0–100) translates behavioral observations into five tiers of collaboration maturity.



**Proficient Tier** • You show proficient AI collaboration skills with solid baseline capabilities. You can plan, evaluate, and iterate effectively with feedback, representing competent collaborative practice.

Figure 3.3 *PAICE score example 1: Proficient*

| Tier | Score Range | Typical Behaviors | Recommended Org Response |
|---|---|---|---|
| **Constrained** | 0–29 | Avoids or misuses AI; frequent errors | Pre-access training required |
| **Informed** | 30–49 | Understands concepts but applies inconsistently | Restricted AI access |
| **Proficient** | 50–69 | Performs reliably with guidance | Standard tool access |
| **Advanced** | 70–89 | Self-corrects and anticipates AI failure modes | Expanded permissions |
| **Exceptional** | 90–100 | Designs systems for human-AI co-creation | Policy input / Center of Excellence roles |

Tier boundaries will continue to calibrate through 2026 pilot partnerships.

Organizations may set their own thresholds based on risk tolerance or regulatory context.

## 3.4 Behavior in Context: Three Illustrative Scenarios

**Marketing Manager (Accountability)** High score: Verifies AI-provided statistics before publication. Low score: Publishes fabricated data; reputational fallout follows.

**Financial Analyst (Integrity)** High score: Rechecks AI's assumptions and documents corrections. Low score: Accepts AI's optimistic forecasts uncritically.

**Developer (Performance + Evolution)** High score: Iterates and learns from AI misfires to improve prompts. Low score: Abandons the tool after one failed attempt.

These patterns show why PAICE measures *behavior* rather than self-reported knowledge: readiness is revealed only when AI is wrong.

**For Leaders:**

- ○ Compare these dimensions to current competency or risk models
- ○ Identify high-variance areas across business units
- ○ Define "AI readiness" goals per function

# 4. How It Works: The PAICE.work Assessment

PAICE.work assessments target real AI collaboration to capture observable behavior, not opinions or recollections.

## 4.1 Adaptive Behavioral Assessment (~20 minutes)

Participants engage in a guided conversation that adapts to their responses.

Those demonstrating strong verification habits progress quickly; others receive scaffolding that exposes how they handle ambiguity and correction.

## 4.2 Strategic Failure Injection

To test Accountability and Integrity, the system introduces subtle factual errors, biased phrasing, or contradictory guidance. This is designed to mimic the same kinds of mistakes AI typically makes, but in an intentional and compressed experience. The participant's reaction (spotting, questioning, or ignoring the anomaly) reveals practical competence.

## 4.3 Real-Time Scoring

The scoring engine evaluates approximately 20 behavioral signals per session.

Algorithms identify verification behavior, contextual judgment, and iteration quality, producing both an overall PAICE Index and five dimension scores.

1. **Access & Consent** – User reads notice and accepts participation terms.
2. **Contextual Scenario** – Role-specific dialogue begins (15–25 minutes).
3. **Behavioral Observation** – The system tracks verification, bias awareness, and iteration quality.
4. **Scoring & Analysis** – Engine maps responses to dimension tiers (~10 seconds).
5. **Feedback & Action Plan** – Personalized insights and team metrics.

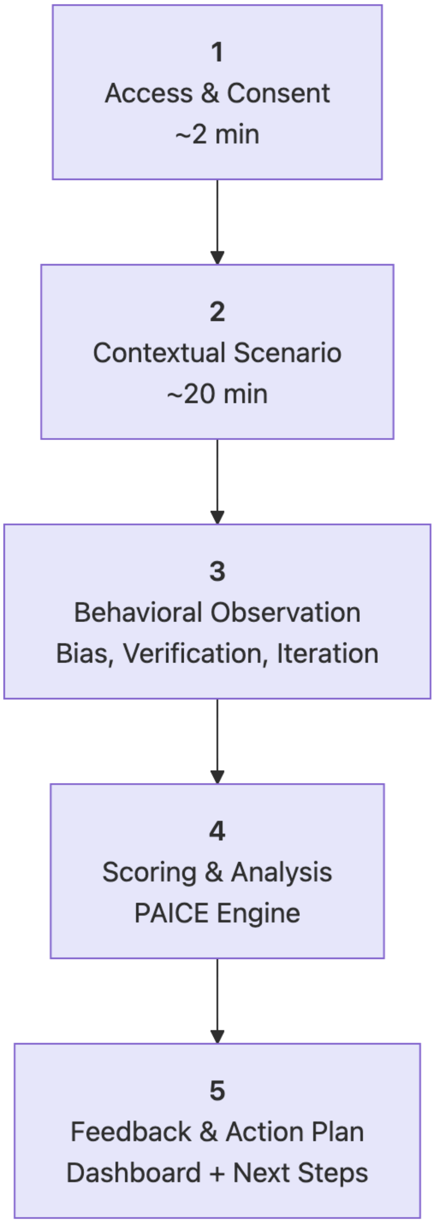


Figure 4.3 *PAICE.work User Journey: Individual*

## 4.4 Dimensional Breakdown and Guidance

Each assessment delivers:

- **PAICE Index (0–100)** – overall collaboration readiness
- **Dimension Scores** – P-A-I-C-E subscores (0–100 each)
- **Tier Classification** – Constrained → Exceptional
- **Targeted Guidance** – development actions aligned to ISO/IEC 42001 and NIST AI RMF

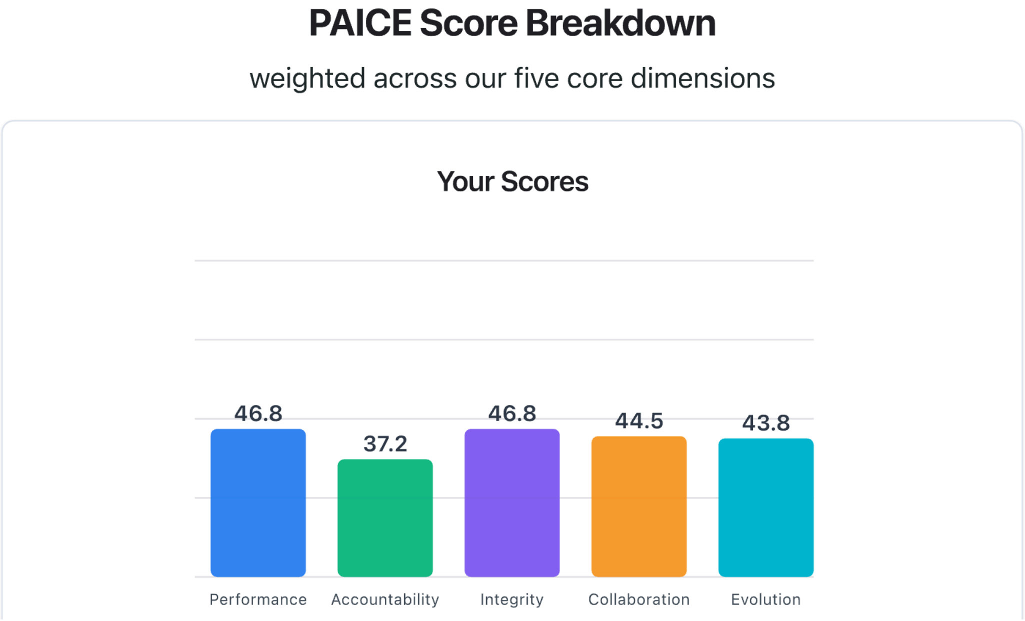


Figure 4.4 *PAICE.work Score Breakdown example 2: Informed*

## 4.5 Designed for Validation

While the behavioral model is functional today, 2026 pilot partnerships will generate the datasets required to confirm:

- Score stability (test-retest reliability)
- Cross-industry generalizability
- Correlation between PAICE tiers and AI-related incident rates

## 4.6 Why PAICE.work is different

| **Traditional Approach** | **Focus** | **Limitation** | **PAICE Advantage** |
|---|---|---|---|
| Knowledge tests | Recall | Measures what people know *about* AI | Observes what they *do with* AI |
| Self-assessments | Confidence | Inflated scores | Objective behavioral data |
| Usage analytics | Activity | Misses failure handling | Captures resilience under stress |
| Training completions | Exposure | No competence proof | Provides quantified capability metrics |

**In short:** PAICE measures the quality of People + AI collaboration when it matters most: *when the system is wrong*.

**For Program Owners:**

- ○ Integrate via secure API with LMS/HRIS
- ○ Use scores to trigger targeted learning plans
- ○ Track score deltas quarterly to monitor AI maturity

*For implementation architecture and scoring logic, see Appendix A: Framework Deep Dive.*

---

# 5. Why It's Defensible: Data Integrity, Security, Accessibility, and Compliance

The following controls describe how the PAICE assessment engine is architected from the ground up for defensibility, pending empirical validation through external research partnerships. Defensibility ensures PAICE can be trusted by design, validation will ensure it can be trusted by evidence.

## 5.1 Anonymization & Data Handling

The PAICE assessment engine is designed according to strict *Privacy by Design* (PbD) principles. All user and session identifiers are one-way hashed with non-reversible salts to prevent re-identification by any means reasonably likely to be used. Conversational inputs are processed transiently in memory and permanently deleted once numeric scores are generated. No textual content, prompts, or interaction logs are retained. The system outputs only anonymized numeric values representing performance across the five PAICE dimensions and ~20 sub-scoring parameters. These scores cannot be linked to individuals or sessions once issued.

Optional contact information, such as an email address supplied by a user who chooses to receive follow-up materials and personalized guidance, is handled under a separate process and stored in an isolated database. It is never connected to any assessment identifier, score, or telemetry data.

On this basis, the PAICE assessment engine operates entirely on **anonymous data** as defined in Recital 26 of the EU General Data Protection Regulation (GDPR) and therefore lies **outside**

**the scope of the GDPR**. The optional contact feature is processed solely under user consent in accordance with Article 6(1)(a).

## 5. 2 Security & Infrastructure Assurance

The PAICE platform follows a defense-in-depth security model consistent with **SOC 2 Type II** and **ISO/IEC 27001** principles. All services are deployed in hardened cloud environments with end-to-end encryption, continuous vulnerability scanning, and role-based access control. Administrative access is logged, reviewed, and periodically audited. Source code and model configurations are version-controlled and integrity-checked through cryptographic hashing.

These controls ensure the PAICE assessment engine meets enterprise-grade standards for confidentiality, integrity, and availability. Security controls are independently reviewable and form part of PAICE's annual risk and compliance audit plan.

## 5.3 Accessibility & Inclusive Design Statement

The PAICE assessment platform follows recognized accessibility and usability standards to ensure equitable participation across roles, regions, and abilities. The system is designed in alignment with the **Web Content Accessibility Guidelines (WCAG) 2.1 Level AA**, emphasizing clarity, keyboard navigation, readable contrast ratios, and screen-reader compatibility.

Assessment uses plain, culturally neutral language and tested to minimize cognitive load, jargon, and idiomatic bias. Alternative formats and reasonable accommodations are supported on request, and accessibility testing is integrated into each major release cycle.

By embedding inclusive design practices into both interface and scenario development, PAICE aims to make the measurement of AI collaboration **accessible, fair, and repeatable** across English-speaking workforces today. Expanded multilingual capability is scheduled for 2026 to extend inclusivity across additional global populations.

## 5.4 Fairness & Bias Mitigation

PAICE scoring models are evaluated through periodic **bias and drift testing** across demographic, linguistic, and industry samples. The system applies a balanced weighting approach to minimize any dimension's disproportionate influence on overall scores. As of 2026, review cycles include external experts from ethics, learning, and data-science domains to monitor representational fairness.

All assessment prompts and data sets undergo pre-release bias screening using both automated detection tools and human review. Results are documented for audit transparency.

### 5.5 Transparency & Explainability

Each PAICE score is accompanied by an interpretable rationale showing which behavioral signals contributed most to the outcome. Weighting logic and dimension definitions are openly documented for enterprise clients and reviewers. This transparency supports internal validation, external audit, and user understanding of how collaborative behaviors influence results. All model updates are accompanied by change logs summarizing the impact on scoring logic and interpretability.

### 5.6 Business Continuity & Version Governance

PAICE maintains version-controlled releases of its framework, scoring models, and data schemas. Each deployment is labeled, documented, and archived to ensure reproducibility of historical scores. Disaster-recovery and backup policies follow **RTO/RPO objectives** consistent with enterprise standards. All historical versions remain queryable to support regulatory look-back or internal audit requests.

**For Risk, Compliance, & Security Teams:**

- ○ Map PAICE logs to internal AI risk documentation standards
- ○ Validate GDPR and localization compliance
- ○ Include PAICE in Responsible AI and cybersecurity audit cycles

---

## 6. Validation & Benchmarking: How PAICE Will Earn Trust

PAICE is operational for individuals today. Enterprise validation now moves from design to evidence. This section defines what exists, what will be measured, and how partners can help establish benchmarks.

### 6.1 What Exists Today

- Individual assessment available at [https://paice.work](https://paice.work)
- Five dimension behavioral framework implemented in production
- Real-time scoring with dimensional breakdowns and tiering
- Adaptive conversational methodology with strategic failure injection
- Personalized, dimension-aligned recommendations

## 6.2 Design Defensibility

- Dimensions align to NIST AI RMF and ISO IEC 42001 control families
- Assessment observes behavior under failure conditions rather than knowledge recall
- Accountability carries the greatest weight because verification is where risk concentrates

## 6.3 Planned Validation Studies

PAICE will run structured studies with partners in 2026. Primary goals are reliability, validity, and practical utility.

**Measurement reliability**

- Test-retest stability across 2 to 6 weeks
- Internal consistency across behavioral signals
- Optional human rater agreement if human scoring modules are added

**Construct validity**

- Convergent and discriminant validity against related capability measures
- Face validity with domain experts in learning, risk, and operations

**Predictive validity**

- Correlation between baseline PAICE tiers and AI related incident rates
- Relationship between Accountability subscore and verification failure frequency
- Impact of targeted training on score deltas and operational outcomes

**Norms and benchmarking**

- Score distributions by role, seniority, and industry
- Cross-organizational ranges for each dimension
- Guidance on risk tolerant thresholds per use case

**Planned metrics**

- Cronbach's alpha internal consistency
- Test retest correlation
- Incident rate ratios by tier
- Effect sizes for pre/post training changes
- Confidence intervals for score interpretation

## 6.4 Early Directional Signals

Unstructured user feedback suggests the format is engaging and the dimensional feedback is actionable. A consistent Accountability gap appears across early users. These signals guide the 2026 validation work but are not presented as proof.

### 6.5 Call for Research Partnerships

PAICE seeks 3 to 5 partners for controlled pilots in Q1-Q2 2026. Partners receive free assessments for defined cohorts, early analytics access, and co-authorship on results. Data sharing is anonymized and limited to scoring outputs and agreed outcome metrics.

**For Analysts:**

  - Compare your internal results to published industry ranges
  - Identify score outliers for intervention or advancement
  - Use score deltas as leading indicators of adoption success

---

## 7. From Assessment to Action: Organizational Integration

PAICE integrates where leaders feel the gap today. The following pilot pathways translate individual scores into organizational controls and learning plans. Each pathway is co validation, not mass deployment.

### 7.1 Pilot Track A. AI Readiness for Rollouts

**Owner**: L&D or HR
**Objective** baseline capability before major AI deployment and measure change after training

- Duration 3 to 6 months
- Cohort 10 to 100 employees across representative roles
- Steps
  - Baseline assessment and role level gap analysis
  - Rollout training and tool access
  - Reassessment at 30 to 90 days
  - Correlate deltas with usage and incident data
- Outputs
  - Readiness report by role and dimension
  - Targeted training plan focused on low dimensions
  - Pre and post evidence for ROI reporting

### 7.2 Pilot Track B. Risk Based Access Controls

**Owner** Risk, Compliance, or InfoSec
**Objective** align tool permissions to demonstrated competence

- Duration 6 to 12 months
- Cohort 20 to 100 employees with mixed seniority
- Steps
    - Baseline assessment and tier classification
    - Implement graduated access policy
    - Monitor incidents and policy exceptions by tier
- Outputs
    - Incident rate analysis by tier
    - Compliance documentation for competence verification
    - Policy guidance on threshold setting

## 7.3 Pilot Track C. Talent Acquisition and Development

**Owner** Talent or HR
**Objective** use behavioral evidence in hiring and internal development

- Duration 6 to 12 months
- Cohort 50 to 200 candidates and employees
- Steps
    - Candidate consented assessments during hiring stages
    - Internal assessments to calibrate talent reviews
    - Longitudinal link to performance data where permitted
- Outputs
    - Talent segmentation by role and tier
    - Predictive validity study against performance indicators
    - Development plans linked to dimension gaps

## 7.4 Integration Patterns

- Learning systems
    - Import results to trigger targeted learning paths
    - Use score deltas for program ROI reporting
- Governance and risk
    - Include PAICE metrics in quarterly AI risk dashboards
    - Map Accountability thresholds to higher risk workflows
- Policy and access
    - Graduated access by tier with documented rationale
    - Reassessment cadence to maintain privileges
- Data handling
    - Scores are anonymized and separable from identity by design
    - Optional email capture is consent based and stored separately

**Implementation note** use first cycle data as baseline, not judgment. Communicate that access and development plans are tied to improvement over time.

**For Implementation Leaders:**

- ○ Begin with one department or region
- ○ Use first-cycle data as baseline
- ○ Share success metrics with AI governance committees

---

## 8. The Broader Mission: Roadmap & Public Benefit

PAICE.work is building shared measurement infrastructure for People + AI collaboration. The roadmap is staged for evidence, scale, and standards adoption.

### 8.1 Phase 1. Pilot Validation

Timeline Q4 2025 to Q2 2026

- Secure 3 to 5 partners across the three tracks
- Baseline 200 to 500 participants
- Publish initial case studies and a predictive validity report
- Establish early benchmark ranges and refine tier thresholds

**Success criteria**

- Demonstrated links between tiers and incident patterns
- Documented training effects on dimension scores
- Two partners progressing to broader deployment

### 8.2 Phase 2. Product Market Fit and Scale

Timeline Q3 2026 to Q4 2026

- Team analytics with dimensional heatmaps and cohort comparisons
- Micro assessments and spaced checks for maintenance
- Enterprise SSO, API for HRIS, and custom weighting by role
- Multilingual support beginning with Spanish, French, Portuguese, and German

**Go to market**

- Design partners shape features
- Case studies drive inbound interest
- Individuals to teams to organizations as a product led path

## 8.3 Phase 3. Infrastructure and Standards

Timeline 2027 to 2028

- Widen organizational benchmarks and publish norms
- Participate in standards work on competence measurement
- Expand to domain specific modules for finance, healthcare, and legal
- Move to multi model scoring and multi modal inputs where appropriate

## 8.4 Public Benefit Commitments

- PBC structure prioritizes mission alignment with commercial sustainability
- Privacy first design with anonymization at capture and minimal retention
- Open data program for research using anonymized datasets under CC BY 4.0
- Transparent limitations and versioned governance for auditability

## 8.5 Risks and Mitigations

- *Predictive validity does not emerge* > Iterate the framework and weighting, pause scale up until resolved
- *Adoption lags* > Extend partner program and focus on team level value delivery
- *Vendor dependence* > Move to multi model ensemble in Phase 2
- *Regulatory shifts* > Maintain mapping to NIST AI RMF and ISO IEC 42001, update guidance as rules evolve

## 8.6 How to Participate

- **Individuals** take the free assessment at https://PAICE.work and get a private baseline
- **Teams** and **Organizations** apply for a pilot through *Pilot [at] PAICE.work* with size, industry, and desired track
- **Researchers** coordinate studies via *Research [at] PAICE.work*
- **Security and compliance teams** request the defensibility packet that includes anonymization, accessibility, security controls, and version governance summaries

**For Stakeholders:**

- Get your score at PAICE.work
- Partner on sector pilots
- Contribute to open research validation

# 9. Acknowledgments & Citations

PAICE exists because of the people who believed in measuring AI collaboration before it was obvious we needed to.

**Special thanks to:**

- **Nate B Jones** and his **Nates's Notes Substack Community** - where this idea evolved from "someone should measure this" to "I'm going to build it"
- **Marily Nika's AI Product Manager cohort** - where the product and the pitch first came together as a viable startup
- **Snapdev.ai team** - for exceptional technical execution and tool support
- **Early testers** (~100+ individuals in Research Preview 2025.10 and 2025.11) - who provided brutal, honest feedback that shaped the framework and revealed the Accountability gap
- **Advisors:**
    - *Barry Kayton* - strategic guidance
    - *Marc Zao-Sanders* - market positioning
    - *Trish Uhl* - AI organizational measurement

**To the future pilot partners:** Thank you for taking a bet on infrastructure-building. Your willingness to share data and co-create evidence will determine whether PAICE.work becomes the standard or remains an interesting research project.

---

## About the Author

**Sam Rogers** is the founder of PAICE.work PBC. Before building PAICE, Sam led global learning technology and analytics initiatives for organizations including YouTube, ADP, Convatec, AAA, and National 4-H Council. His work spans AI adoption strategy, risk-aware implementation, and systems governance across complex enterprise environments.

A long-time advocate for measurable learning and responsible AI and frequent speaker, he created the PAICE framework to define and benchmark how people and AI collaborate effectively.

**Contact:** srogers [at] PAICE.work
**LinkedIn**: https://www.linkedin.com/in/samrogers/

## About the Company

**PAICE.work** is a Public Benefit Corporation, structured to balance impact and returns while prioritizing the mission "to enable safer and more effective People+AI collaboration by providing independent capability measurement."

We are currently pre-seed and pre-revenue, with no external funding as of November 2025. Soon we will begin seeking strategic investors aligned with responsible AI and long-term infrastructure plays.

**Contact:** https://paice.work/contact
**LinkedIn**: https://www.linkedin.com/company/paice-work/

---

## Citation & License

**Citation:**
Rogers, S. (2025). *PAICE.work: Making AI Collaboration Measurable, Teachable, and Governable - Vision & Partnership Whitepaper (v3.1)*. PAICE.work PBC. https://PAICE.work/whitepaper

**License:**

*This is Research Preview 2025.11. PAICE is operational for individual assessment and seeking pilot partners for organizational validation. Claims about organizational impact are projections based on framework design, not proven outcomes. See Appendix E for detailed limitations.*

# APPENDICES

## Appendix A: Framework Deep Dive

### Full Dimensional Definitions

This appendix provides technical detail on the five PAICE dimensions, scoring methodology, and tier system. Though specific sub-scoring parameters remain proprietary, approved researchers can access them post-NDA. For the business case and pilot pathways, see main body sections 3-6.

### Scoring Calibration Reference

PAICE scores are calibrated to represent a meaningful spectrum of collaboration capability:

- **0** = Actively hostile or non-cooperative with AI systems (refuses to engage, sabotages outputs)
- **100** = World-class AI collaboration (equivalent to Andrej Karpathy on his best day)

Most people fall in the middle ranges (30-70), which accurately reflects current workforce capability. This calibration ensures scores are grounded in real-world performance rather than arbitrary percentage cutoffs.

### Dimension 1: Performance

**Definition:** The ability to formulate clear tasks, craft effective prompts, and use AI efficiently to achieve work objectives.

**What we measure:**

- **Task clarity:** Can the user articulate what they want from AI with sufficient specificity?
- **Prompt quality:** Do prompts include necessary context, constraints, and success criteria?
- **Iteration efficiency:** How many rounds does it take to get usable output?

**Behavioral indicators of high Performance:**

- Prompts include role, context, format, and constraints upfront
- User recognizes "good enough" vs. over-iterating
- Efficient use of AI features (e.g., Claude Projects for context, GPT custom instructions)
- Ability to break complex tasks into AI-manageable subtasks

**Behavioral indicators of low Performance:**

- Vague prompts requiring AI to guess intent
- Over-reliance on default settings without customization
- Inability to evaluate output quality (accepts anything AI produces)
- Frustration loops (repeatedly failing to get usable output)

**Why it matters:**
Low Performance means wasted time and poor AI ROI. Even if Accountability is high (verification is strong), if Performance is low, AI adds friction rather than value.

## Dimension 2: Accountability

**Definition:** The capability to detect errors, verify outputs, and maintain responsibility for AI-assisted work.

**What we measure:**

- **Error detection:** Does the user catch hallucinations, factual errors, logical inconsistencies?
- **Bias awareness:** Can they identify potentially biased or unfair AI recommendations?
- **Output ownership:** Do they treat AI outputs as drafts requiring validation, or as final work?

**Behavioral indicators of high Accountability:**

- Systematic verification of AI claims against authoritative sources
- Recognition of hallucination patterns (e.g., fabricated citations, confident falsehoods)
- Appropriate skepticism calibrated to task risk (higher scrutiny for legal/medical vs. brainstorming)
- Clear documentation of "AI-assisted" vs. "human-verified" in outputs

**Behavioral indicators of low Accountability:**

- Accepting AI outputs without verification
- Treating confidence of AI tone as signal of accuracy (it's not)
- Failure to recognize obvious errors (dates, calculations, contradictions)
- Publishing AI outputs as-is in high-stakes contexts

**Why it matters:**
Accountability failures create organizational liability. The legal brief with fabricated citations, the biased hiring recommendation, the confidential data leak—these are Accountability dimension failures.

**This is the highest-weighted dimension in PAICE Index scoring** (30% of total), reflecting that verification failures cause the most acute organizational harm. Early data confirms this is also the consistently lowest-scoring dimension across users, revealing the critical capability gap.

## Dimension 3: Integrity

**Definition:** Ensuring AI-assisted outputs remain factually grounded, logically consistent, and aligned with truth.

**What we measure:**

- **Factual grounding:** Does the user ensure claims are supported by evidence?
- **Logical consistency:** Do they catch internal contradictions in AI reasoning?
- **Ethical guardrails:** Does the user reject outputs that are technically accurate but ethically problematic?

**Behavioral indicators of high Integrity:**

- Cross-checking AI-generated facts against primary sources
- Identifying logical fallacies or circular reasoning in AI outputs
- Refusing to use AI for deceptive purposes (even if technically possible)
- Transparency about limitations ("AI suggests X, but confidence is low because Y")

**Behavioral indicators of low Integrity:**

- Accepting AI outputs without concern for truth
- Using AI to generate plausible-sounding falsehoods
- Failure to identify contradictions between AI response and known facts
- Prioritizing AI convenience over accuracy

**Why it matters:**
Integrity separates "efficient but wrong" from "efficient and trustworthy." Low Integrity scores indicate users who will amplify AI errors rather than correct them.

**Standards Alignment:**

- ISO/IEC 42001:2023 Control 5.2 (Risk Assessment)
- NIST AI RMF GOVERN 1.1 (Legal and regulatory requirements)
- EU AI Act Article 14 (Human oversight)

## Dimension 4: Collaboration

**Definition:** The quality of People + AI interaction workflow, including iteration, feedback, and strategic use of AI capabilities.

**What we measure:**

- **Iteration quality:** Do refinements improve output, or just spin wheels?
- **Feedback specificity:** Can the user articulate *why* output is insufficient and *how* to improve it?
- **Workflow integration:** Is AI use fluid, or does it create friction in existing processes?

**Behavioral indicators of high Collaboration:**

- Structured iteration ("This output is good for X, but lacks Y—please add Z")
- Knowing when AI adds value vs. when it's overkill
- Seamless handoffs between AI drafting and human refinement
- Using AI as thought partner, not just output generator

**Behavioral indicators of low Collaboration:**

- Vague feedback loops ("make it better")
- Using AI for tasks where manual work would be faster/better
- Treating AI as purely transactional (prompt → output, no iteration)
- Workflow disruption (AI use slows down rather than accelerates work)

**Why it matters:**
Collaboration quality determines whether AI feels like a productivity multiplier or an annoying chatbot. High Collaboration scores mean fluid People + AI partnership. Low scores mean friction.

**Standards Alignment:**

- ISO/IEC 42001:2023 Control 7.2 (Competence)
- NIST AI RMF MANAGE 2.3 (Mechanisms for operator override)

## Dimension 5: Evolution

**Definition:** The capacity to learn from AI interactions, recognize patterns, and adapt collaboration strategies over time.

**What we measure:**

- **Reflection:** Does the user analyze what worked/didn't work in AI interactions?
- **Skill development:** Are they improving AI collaboration over time, or stuck in habits?
- **Adaptation:** Do they update their approach when AI capabilities change?

**Behavioral indicators of high Evolution:**

- Active reflection on AI interaction quality ("This prompt worked better than last time because...")
- Building mental models of AI strengths/weaknesses
- Experimenting with new prompting techniques or AI features
- Adjusting workflow as AI tools improve

**Behavioral indicators of low Evolution:**

- Repeating ineffective prompting patterns without learning
- No awareness of what works vs. what doesn't
- Treating AI as static tool (no adaptation as capabilities evolve)
- Frustration without diagnosis ("AI is bad" vs. "I need to change my approach")

**Why it matters:**
Evolution separates users who plateau vs. those who compound AI capability over time. As AI systems improve (monthly model updates), high-Evolution users extract increasing value. Low-Evolution users stagnate.

**Standards Alignment:**

- ISO/IEC 42001:2023 Control 10.2 (Continual improvement)
- NIST AI RMF MANAGE 4.1 (Monitoring and review)

## Standards Mapping Table

| PAICE Dimension | ISO/IEC 42001:2023 | NIST AI RMF | EU AI Act |
|---|---|---|---|
| **Performance** | Control 7.2 (Competence)<br>Control 8.1 (Operational planning) | MANAGE 1.1 (Allocation of resources)<br>MANAGE 2.1 (Planning) | Article 9 (Risk management system) |
| **Accountability** | Control 5.2 (Risk assessment)<br>Control 9.1 (Performance evaluation) | GOVERN 1.1 (Legal requirements)<br>MANAGE 2.2 (Accountability) | Article 14 (Human oversight)<br>Article 17 (Quality management) |
| **Integrity** | Control 6.2 (Information security)<br>Control 8.2 (Testing) | GOVERN 1.2 (Ethical guidelines)<br>MAP 1.1 (Context) | Article 10 (Data governance)<br>Article 15 (Accuracy) |
| **Collaboration** | Control 7.3 (Awareness)<br>Control 8.3 (Communication) | MANAGE 2.3 (Operator override)<br>MANAGE 3.1 (Documentation) | Article 13 (Transparency)<br>Article 14 (Human oversight) |
| **Evolution** | Control 10.2 (Continual improvement)<br>Control 9.2 (Monitoring) | MANAGE 4.1 (Monitoring)<br>MANAGE 4.2 (Performance metrics) | Article 72 (Post-market monitoring) |

This mapping provides compliance teams with specific controls/functions to reference when demonstrating AI collaboration competence to auditors.

## Scoring Methodology

**PAICE Index Calculation:**

Each dimension receives a 0-100 score based on ~20 behavioral parameters. The PAICE Index is a weighted average:

- **Performance:** 10% weight
- **Accountability:** 30% weight (highest, verification failures cause acute harm)
- **Integrity:** 20% weight
- **Collaboration:** 25% weight
- **Evolution:** 15% weight

**Tier Classification:**

| PAICE Index | Tier | Description |
|---|---|---|
| 0-29 | Constrained | Struggles to integrate or critique AI output; frequent misalignment |
| 30-49 | Informed | Conceptually aware of AI behavior but reactive or rigid in practice |
| 50-69 | Proficient | Solid baseline; can plan, evaluate, and iterate effectively with feedback |
| 70-89 | Advanced | Self-correcting, anticipates AI failure modes, reliable collaborator |
| 90-100 | Exceptional | Emergent fluency; human–AI co-creation at system-level alignment |

**Note:** Tier boundaries are calibrated based on Research Preview data and will be refined through pilot partnerships. Organizations may set custom thresholds (e.g., "Our high-risk use cases require Proficient+ tier").

## Adaptive Difficulty

PAICE assessments adjust difficulty based on user performance:

- **Strong performers** bring harder scenarios and tend to finish faster (~15 min)
- **Struggling users** receive more scaffolding and guidance (~25 min)
- **Adaptive responses** ensure scores reflect capability, not just exposure to AI or knowledge about how it works

This prevents "teaching to the test" gaming, users can't improve scores by memorizing content, only by collaborating better with AI.

## Strategic Failure Injection

To measure Accountability effectively, PAICE systematically introduces errors into AI responses:

**Subtle failures:**

- Plausible but incorrect statistics
- Outdated information presented as current
- Logically consistent but factually wrong claims
- Biased language that sounds neutral

**Obvious failures:**

- Internal contradictions
- Impossible dates or numbers
- Fabricated sources
- Clear logical fallacies

The mix of subtle and obvious failures reveals whether users:

- Catch only blatant errors (low Accountability)
- Detect nuanced problems (high Accountability)
- Verify systematically vs. spot-check randomly
- Maintain appropriate skepticism under time pressure

This methodology was inspired by research on human error detection and red-teaming practices in AI safety.

### Anti-Gaming Measures

Preventing score manipulation:

1. **User scenario driven:** Users bring their own real-world situation to the assessment, there is no set pattern or scriptable scenario bank to chose from
2. **Behavioral pattern analysis:** We detect scripted responses, copied prompts, or AI-assisted assessment-taking
3. **Confidence calibration checks:** Overconfident users who fail verification tasks are scored down more than uncertain users who verify correctly
4. **Diminishing returns on iteration:** Excessive prompt refinement without improvement signals low collaboration quality
5. **Agentic browser detection**: Prevents automated assessment-taking[1]

**Consequence of detected gaming:** Score adjusted downward + flagged for manual review. Strong disincentive to cheat.

---

## Appendix B: Assessment Design

### Conversational Evaluation Methodology

Unlike traditional multiple-choice tests, PAICE uses a conversational interface where users interact with AI systems in their own real-world scenarios. The assessment observes *behavior*, not just knowledge recall.

**Example scenario (Accountability dimension):**

> **PAICE:** "You asked me to summarize a research paper on climate change impacts. Here's the output:
>
> *'According to a 2023 study published in Nature, global temperatures have risen 1.5°C since pre-industrial times, primarily due to fossil fuel emissions. The study, authored by Dr. Sarah Chen and colleagues, projects 2.7°C warming by 2100 under current policies. Notably, the IPCC's 2022 report supports these findings.'*
>
> This is now ready for your client report."

[1] Technical details on agentic browser detection and security measures available in blog post: [Protecting PAICE: Our Agentic Browser Detection Strategy]

**User response analysis:**

- **High Accountability:** "I'd like to verify the Nature citation (author, year, title), cross-check the 1.5°C figure against IPCC AR6, confirm the 2.7°C projection source, and validate that Chen et al. actually exists and published on this topic."
- **Medium Accountability:** "Check if the IPCC report agrees with the temperature figures."
- **Low Accountability:** "Thanks! Next..." or "This is correct, right?"

The assessment scores not just the answer, but the *behavioral indicators*:

- Did they identify specific claims to verify?
- Did they name appropriate verification sources?
- Did they recognize the risk (client report = high stakes)?
- Did they demonstrate systematic verification workflow?

## Data Privacy & Retention

**During assessment:**

- Conversation is analyzed in real-time for behavioral patterns
- Text is processed but not stored long-term
- No identifiable user data unless explicitly provided

**After assessment:**

- **Conversation text:** Never captured by PAICE.work architecture
- **Behavioral vectors:** Retained (anonymized patterns, no raw text)
- **PAICE scores:** Stored with user consent (can be deleted anytime)

**Third-party processing:**

- Conversations are processed through Anthropic's Claude API during assessment
- Anthropic may retain data according to their Commercial Terms (currently deleted after 30 days)
- PAICE does not control Anthropic's data practices
- We chose Anthropic specifically for their commitment to privacy and responsible AI

**User rights:**

- View all stored data
- Export data in portable format
- Opt out of research use

**Organizational data:**

- Aggregated scores (team/cohort level)
- Incident correlation data anonymized before sharing with PAICE
- Pilot partners sign data sharing agreements specifying retention/usage

**Compliance:**

- GDPR-compliant (right to deletion, portability, access)
- CCPA-compliant (California privacy rights)
- SOC 2 Type II audit planned for 2026

For complete privacy details, see PAICE.work/privacy or our more user-friendly blog post Your Data, Your Privacy: How PAICE Handles Your Information

## Security Practices

**Assessment integrity protection:**

- Agentic browser detection to prevent automated assessment-taking[2]
- Behavioral pattern analysis to detect gaming attempts
- Environment-aware enforcement (development/staging/production)
- Privacy-first honeypot redirects (non-confrontational approach)

**Data protection:**

- Encrypted data transmission (HTTPS)
- Secure data storage
- Access controls and authentication
- Regular security reviews
- Limited data retention

**No surveillance:**

- Individual users control their data
- Results are not shared without consent
- Assessment is voluntary
- Focus is on learning, not punishment

# Appendix C: Organizational Implementation Guide

This appendix provides detailed guidance for organizations deploying PAICE across teams. For pilot partnership tracks, see *Section 6* in main body.

## Pre-Deployment: Building the Case

**Step 1: Identify the business problem**

[2] Technical details on agentic browser detection and security measures available in blog post: Protecting PAICE: Our Agentic Browser Detection Strategy

Which of these resonates most strongly?

- ⚠️ **Risk exposure:** "We don't know who's using AI safely vs. creating liability"
- 💰 **Wasted L&D spend:** "We trained everyone, but can't prove it worked"
- 🚫 **One-size-fits-none policy:** "High performers are constrained by policies for low performers"
- 📋 **Compliance gap:** "Auditors want competence evidence beyond *training completed*"

Your answer determines messaging, stakeholder buy-in strategy, and success metrics.

**Step 2: Secure executive sponsorship**

PAICE deployments succeed when sponsored by:

- **CHRO / VP People:** If framing as talent development / capability measurement
- **CRO / Chief Risk Officer:** If framing as risk mitigation / competence verification
- **CIO / CTO:** If framing as AI enablement / graduated access controls
- **CLO / VP Learning:** If framing as training ROI / evidence-based development

**Pitch template:**

> "We're deploying AI tools to [X employees] but can't answer 'are they ready?' Training completion rates don't measure competence, they measure exposure. PAICE measures what people *do* when AI fails, giving us:
>
> - **Risk quantification:** Baseline capability before rollout
> - **Targeted development:** Where to focus L&D resources
> - **Compliance evidence:** Competence verification auditors will accept
> - **Graduated access:** Safe AI empowerment for high performers
>
> Cost: [free for pilot, $X per user for enterprise - TBD] Timeline: 3-6 months to baseline + validate ROI: [reduced incidents, focused training spend, defensible competence program]"

**Step 3: Define success metrics**

What will prove PAICE.work worked? Examples:

- **Risk reduction:** "AI-related incidents decrease 30% in high-scoring cohorts"
- **Training ROI:** "Post-training PAICE scores improve 15+ points + operational improvement"
- **Policy effectiveness:** "Graduated access based on tiers reduces incidents without productivity loss"
- **Compliance value:** "Auditors accept PAICE scores as competence evidence"

Always write success criteria *before* deployment to avoid moving goalposts.

## Phase 1: Baseline Assessment (Weeks 1-2)

**Goal:** Understand current state before AI rollout or intervention

**Cohort selection:**

- **Representative sample:** 10-20% of eventual AI user population
- **Cross-functional:** Include multiple roles, seniority levels, departments
- **Voluntary initially:** Mandatory assessment creates resistance; start with volunteers

**Recommended cohort size:**

- Small team (5-10 AI users): Assess everyone
- Mid-size (11-25 users): Assess CoE & designated representative users
- Enterprise (500+ users): Assess 50-100 across business units

**Logistics:**

- Send email from executive sponsor explaining why ("we're measuring AI readiness")
- Provide PAICE link + estimated time (15-25 min)
- Emphasize: no preparation needed, authentic behavior produces best results
- 2-week window for completion
- Reminder at 1 week, escalation path for non-responders

**Messaging tips:**

- ✅ "This helps us understand where to focus training resources"
- ✅ "Your score is for development, not performance review"
- ✅ "We're measuring organizational readiness, not individual ranking"
- ❌ "Everyone must complete by Friday" (creates compliance burden)
- ❌ "Low scores will restrict AI access" (save graduated access for Phase 3)

**Data collection:**

- PAICE Index + dimensional scores for cohort
- Role, department, seniority (for segmentation)
- Prior AI experience (self-reported)

**Analysis:**

- Mean/median PAICE scores by dimension
- Distribution across tiers (what % are Proficient+?)
- Gaps by role (which teams need most support?)
- Priority dimensions (where are org-wide weaknesses?)
- Accountability gap magnitude (expected: 10-20 points lower than average)

**Baseline report template:**

**Organizational AI Readiness - Baseline Assessment**

**Cohort:** 47 employees across Sales, Finance, Operations
**Assessment period:** Jan 1-15, 2026
**Participation rate:** 89% (42/47 completed)

**Overall Results:**

- Mean PAICE Index: 54 (Proficient tier)
- Tier distribution: 12% Constrained, 31% Informed, 45% Proficient, 10% Advanced, 2% Exceptional

**Dimensional Breakdown:**

- Performance: 62 (strongest dimension)
- Accountability: 48 (organizational gap—verification weak)
- Integrity: 56
- Collaboration: 52
- Evolution: 51

**Key Findings:**

- 43% of cohort is below Proficient tier → high-risk for deployment without support
- Accountability is lowest dimension (consistent with research data) → focus training on verification workflows
- Sales team scores 10 points lower than Finance → role-specific needs

**Recommendations:**

- Deploy Accountability-focused training before AI rollout
- Consider graduated access (Informed tier = restricted, Proficient+ = full access)
- Reassess in 60 days post-training to measure improvement

## Phase 2: Targeted Development (Months 1-3)

**Goal:** Improve capability in identified gap areas

**Intervention options:**

1. **Dimension-aligned training**
   a. If Performance is low: Task decomposition, output evaluation
   b. If Accountability is low: Verification workflows, fact-checking training
   c. If Integrity is low: Ethical AI use, bias awareness
   d. If Collaboration is low: Prompt engineering, iteration techniques
   e. If Evolution is low: Reflective practice, experimentation frameworks

2. **Role-specific workshops**
    a. Sales: AI for prospecting, competitive intel (focus on Integrity)
    b. Finance: AI for analysis, forecasting (focus on Accountability)
    c. Operations: AI for process optimization (focus on Collaboration)
3. **Mentorship pairing**
    a. High scorers (Advanced/Exceptional tier) mentor low scorers
    b. Peer learning often more effective than formal training
4. **Just-in-time resources**
    a. Slack bot with PAICE tips ("Today's tip: Always verify AI citations")
    b. Email series on dimensional improvement
    c. Lunch-and-learn sessions

**Tracking engagement:**

- Who completed training?
- Time spent in learning resources
- Self-reported behavior change ("I now verify AI outputs before sharing")

**Avoid:**

- Generic "intro to AI" training. PAICE identifies *specific* gaps, target those.
- Mandatory training without explaining why. This creates resentment.
- Training without pre/post-assessment. Can't prove it worked.

### Phase 3: Proof of ROI (Quarter 1)

**Goal:** Demonstrate capability improvement + link to operational metrics

**Re-assessment:**

- Same cohort, 30-90 days post-training
- Compare pre/post PAICE scores
- Expected improvement: 10-20 points on targeted dimension

**Operational metrics to track:**

- **Incidents:** AI-related errors, policy violations (decreased?)
- **Productivity:** Time to complete AI-assisted tasks (increased?)
- **Quality:** Peer review scores on AI-assisted work (improved?)
- **Compliance:** Audit findings related to AI use (reduced?)

**ROI calculation framework:**[3]

**Example:**

- Training cost: $50/user × 50 users = $2,500
- PAICE assessment cost: Free for pilot (later: $10/user = $500)
- Total investment: $3,000

**Measurable outcomes:**

- AI-related incident reduction: 5 incidents → 1 incident (estimated cost of incident: $10,000) = $40,000 avoided
- Training efficiency: Focused on Accountability gap (30% of training time) vs. generic training (100% of training time) = $5,000 saved
- Compliance value: Auditors accept PAICE scores, reducing audit time by 20 hours ($200/hr) = $4,000 saved

**Total ROI:** $49,000 in value / $3,000 investment = **16x return**

*Note: These are illustrative figures. Actual ROI depends on your incident costs, training expenses, and operational context.*

**Case study template:**

**Improving AI Collaboration Accountability at [Company]**

**Challenge:** Deploying Microsoft 365 Copilot to 500 employees without baseline competence measurement

**Solution:** PAICE baseline assessment (N=50) identified Accountability as lowest dimension (mean: 48)

**Intervention:** Targeted verification training focused on:

- Fact-checking workflows
- Hallucination detection
- Citation validation

**Results:**

- Post-training Accountability scores improved from 48 → 63 (+15 points)
- AI-related errors decreased 40% in trained cohort vs. untrained
- Training resources focused on actual gap (30% time savings vs. generic training)

[3] Full ROI measurement framework with case studies available in blog series: [Measuring AI Collaboration ROI, Part 1: Framework and Metrics]

> **Key Insight:** Behavioral measurement revealed Accountability gap that usage metrics missed. Targeted training delivered measurable improvement.

## Phase 4: Scale (Quarter 2+)

**Goal:** Expand beyond pilot cohort to full workforce

**Rollout strategy:**

**Option A: Gradual expansion**

- Quarter 1: Pilot cohort (~50 users)
- Quarter 2: Expand to department (~200 users)
- Quarter 3: Full rollout (all AI tool users)

**Option B: Graduated access triggering**

- All employees assess before AI tool access
- Tier determines initial permissions:
  - Constrained: Prerequisite training required before access
  - Informed: Restricted access (supervised use, limited features)
  - Proficient+: Full access

**Option C: Hiring integration**

- Add PAICE to interview process (with candidate consent)
- Use scores to inform AI-adjacent role hiring
- Track correlation with performance reviews over time

**Organizational features (2026 launch):**

- **Team dashboards:** Aggregated PAICE scores, dimensional heatmaps
- **Trend tracking:** Score changes over time (quarterly re-assessment)
- **Cohort comparison:** How does Sales compare to Finance?
- **Training targeting:** Auto-generate recommended training based on gaps
- **MCP integration:** Model Context Protocol for embedded analytics, dedicated assessments no longer needed once PAICE included in AI workflow

## Common Implementation Mistakes to Avoid

Drawing from research and early pilot conversations, here are mistakes organizations should avoid:[4]

[4] Complete catalog of mistakes and solutions available in blog post: Common AI Collaboration Mistakes (And How to Avoid Them)

**1. The "Copy-Paste-Submit" Culture**

- **Mistake:** Not establishing verification standards, allowing users to publish AI outputs without review
- **Solution:** Make verification non-negotiable for high-stakes outputs, train on what "good enough" verification looks like

**2. Vague Deployment Mandates**

- **Mistake:** "Everyone must use AI" without clarity on what good use looks like
- **Solution:** Use PAICE tiers to set expectations ("Proficient+ users get advanced features")

**3. Ignoring the Accountability Gap**

- **Mistake:** Assuming users will naturally verify AI outputs
- **Solution:** Accept that Accountability is hardest skill, allocate 2x training resources to verification

**4. Treating PAICE as Performance Review**

- **Mistake:** Using scores punitively before people have development opportunity
- **Solution:** Frame as development tool first, only integrate into performance after baseline + training

**5. Over-Reliance on Single Metric**

- **Mistake:** Using only PAICE Index, ignoring dimensional breakdown
- **Solution:** Analyze dimension-specific gaps, target training accordingly

**6. No Systematic Follow-Up**

- **Mistake:** Assess once, never reassess or track improvement
- **Solution:** Quarterly reassessment for active learners, annual for maintenance

For complete catalog of AI collaboration mistakes, see Common AI Collaboration Mistakes.

## Change Management: Making Assessment Non-Threatening

**Common resistance points**

1. **"This feels like surveillance"**
   *Response*: "PAICE measures capability for development, not surveillance. Conversations are confidential to you, scores are measured to show improvement. Progress and general strengths and weaknesses across your entire team is what is shared with managers."
2. **"I don't have time for another assessment"**
   *Response*: "15-25 minutes now prevents hours of AI-related rework later. Plus, it's actually useful, you'll learn where to focus your own development."

3. **"What if I score low?"**

   *Response*: "Low scores aren't failures, they're baselines. Most people are in the Informed or Proficient tiers, and Accountability is consistently lowest dimension. The goal is improvement, not judgment."
4. **"My job doesn't use AI"**

   *Response*: "Yet. AI is becoming infrastructure. Measuring readiness *before* deployment is exactly the point."

**Best practices**

- ✅ **Voluntary before mandatory:** Start with volunteers, build social proof, then expand

  ✅ **Executive participation:** Leadership takes PAICE first, they choose to share their scores (if comfortable)
- ✅ **Privacy commitments:** Scores not used for performance reviews initially (save that for Phase 4)
- ✅ **Actionable feedback:** Every user gets specific and personalized development recommendations, not just a number
- ✅ **Celebrate improvement:** Recognize users who improve scores, not just high scorers
- ❌ **Avoid:** Launching assessment with AI access restrictions, this creates fear
- ❌ **Avoid:** Ranking employees by PAICE score publicly, this creates competition, not learning

  ❌ **Avoid:** Assessment without explanation of why, this creates resentment & confusion

## Support Resources for Deployment

**Interpretation guides**

- What each dimension means in plain language
- How to read PAICE reports
- What tier classifications indicate

**Learning libraries**

- Map dimensional gaps to training resources (internal or external)
- Curated resources for each dimension:
  - Accountability: Verification workflows, fact-checking tools
  - Integrity: Ethical AI use, bias awareness
  - Collaboration: Prompt engineering courses
  - Performance: Task decomposition frameworks
  - Evolution: Reflective practice guides

**Community of practice**

- Slack channel or internal forum for PAICE participants
- Peer learning ("How did you improve your Collaboration score?")
- Q&A with L&D team

**Manager toolkit**

- How to discuss PAICE scores with direct reports
- Coaching conversation templates
- Development planning guides

## Common Pitfalls & Solutions

| **Pitfall** | **Cause** | **Solution** |
|---|---|---|
| **Low engagement** | Unclear value proposition | Executive sponsorship + clear "why this matters" messaging |
| **Disappointing scores** | Unrealistic expectations (expecting high scores) | Normalize baselines ("most people are Informed tier initially"), celebrate growth |
| **Resistance to "AI testing"** | Surveillance fears | Emphasize development not evaluation, privacy commitments, voluntary initial phase |
| **Score misuse as ranking** | Manager misunderstanding | Training on proper interpretation, non-punitive culture, clear policy on score use |
| **Limited budget for training** | Assessment without follow-up action | Start with high-leverage gaps (Accountability), phased training rollout, peer mentorship (free) |
| **Assessment fatigue** | Too frequent re-assessment | Quarterly max for individuals, annual for low-stakes roles, only re-assess if intervention occurred |

# Appendix D: Frequently Asked Questions

## About the Assessment

**Q: How long does an assessment take?**
A: Between 15-25 minutes, 20 minutes on average. For best results, dedicate 30 minutes of uninterrupted time and complete your session in one sitting.

**Q: Should I prepare for the assessment?**
A: No. PAICE measures natural collaboration behavior, not test-taking ability. Preparation actually reduces accuracy because we want to see your authentic patterns, not rehearsed behaviors. Just bring a real work task and work naturally.[5]

**Q: Can I pause and resume?**
A: Yes. Your session can be paused and continued later, however it is not advised. This is not due to any system constraints, it just makes it more difficult for people to score their best this way.

**Q: Can users game the system?**
A: Not easily. Weighting and user scenarios vary per session. Confidence checks detect scripted behavior. Gaming attempts are flagged and scores adjusted downward, a strong disincentive.

**Q: Will you publish the system prompts or scoring algorithms?**
A: No. Scoring algorithms remain proprietary to preserve measurement integrity. However, dimensional definitions and framework rationale are transparently documented in this whitepaper.

**Q: Is this an AI assessment or a human assessment?**
A: Both. PAICE measures the *interaction quality*, or how humans and AI perform together. It's not testing AI capabilities or human capabilities in isolation.

**Q: Why do you inject failures into AI responses?**
A: To measure Accountability, the ability to detect and recover from AI errors. This is the highest-weighted dimension because failure detection is the most critical skill for reducing organizational risk.

## About Data & Privacy

**Q: What happens to my data after assessment?**
A: Conversation text is never captured in PAICE.work production systems. Only anonymized behavioral vectors (scoring patterns, not raw text) are retained, and these are stored with your hashed user identity (not personally identifiable).

**Q: Can my employer see my individual PAICE score?** A: Only if you consent. For pilot partnerships, scores are aggregated (team-level averages) unless you opt to share more. After pilot phase, enterprise deployments will have clear data sharing policies aligned to specific organizational needs.

[5] Comprehensive preparation guidance available in blog post: How to Prepare for Your PAICE Assessment (Spoiler: You Don't)

**Q: Is PAICE GDPR/CCPA compliant?**
A: Yes. You have rights to access, delete, and export your data. We minimize data collection and retention. See Privacy & Data Practices for complete details.

**Q: How do you prevent bias in scoring?**
A: Ongoing bias audits across demographic groups (when self-reported). Dimensional framework designed to avoid culture-specific knowledge requirements. If bias is detected, we recalibrate. Quarterly transparency reports will document validation findings.

**Q: Who processes my conversation during assessment?**
A: Your conversation is processed through Anthropic's Claude API. Anthropic may retain data according to their Commercial Terms. We chose Anthropic for their commitment to privacy and responsible AI. In the future, we will leverage other AI models as well (ChatGPT, Gemini, etc.) and we will update our third-party policies with transparency at PAICE.work/privacy.

## About Scoring & Tiers

**Q: What's a "good" PAICE score?**
A: Most people fall in Informed (30-49) or Proficient (50-69) tiers. "Good" depends on context:

- For individual development: Any score gives you a baseline to improve from
- For organizational risk: Proficient+ (50+) is generally safe for standard AI access
- For high-stakes use cases: Advanced+ (70+) may be required

The median score is around 42 (Informed tier), accurately reflecting current workforce capability.

**Q: Why is my Accountability score so low?**
A: This is the most common pattern. Accountability (error detection, verification) is consistently the lowest-scoring dimension across users, most people score 10-20 points lower on Accountability than their average. This isn't a personal failing, it's the hardest skill to master. See Why Your Accountability Score Is Lower for detailed explanation.

**Q: How often should I/we re-assess?**
A: As often as you like, but not sooner than 7 days. We generally recommend:

- **Individuals:** *Every 4-6 weeks during active AI skill development.* We will also offer a paid product that is designed to supercharge AI skill development.
- **Teams:** *Monthly or quarterly during AI rollouts, annually for maintenance.* We will also offer a cohort micro-assessment product that measures specific sub-scores in shorter sessions (5-10 minutes) more frequently over time.
- **Organizations:** *Annually for compliance, quarterly for transformation initiatives.* We will also offer an enterprise-embedding product that doesn't require any dedicated assessment time from users.

**Q: What if someone scores in Constrained tier (below 30)?**
A: This signals important foundational gaps in AI collaboration capability (or an attempt at gaming the assessment), it does *not* indicate lack of intelligence. Recommendation: Prerequisite training on basic AI concepts, verification practices, and ethical awareness before granting

access to production AI tools. *Constrained-tier users remain high-risk without intervention and support.*

**Q: Can PAICE scores improve?**
A: Yes. Scores typically improve with:

- Targeted training on low-scoring dimensions
- Prolonged AI usage with reflective practice
- Mentorship from high-scoring users

Expected improvement: 10-20 points on targeted dimension over 30-90 days with intervention.

## About Organizational Deployment

**Q: Do you support languages other than English?**
A: Not yet. Multilingual support (Spanish, French, Portuguese, German) is planned for 2026.

**Q: How does PAICE handle neurodiverse users or accessibility needs?**
A: Our assessment (and our entire website) are:

- WCAG 2.1 AA compliant
- screen reader compatible
- keyboard navigable

Conversational format reduces barriers vs. traditional tests. Extended time accommodations available upon request. *We believe AI collaboration is for everyone*, which is why we designed PAICE.work to be accessible from the beginning. We actively seek feedback on accessibility improvements. If you have suggestions or unique needs that we can better accommodate, please reach out to us at PAICE.work/contact.

**Q: Can we customize dimensional weighting for our use cases?**
A: Not yet, but this is on our roadmap (see *Section 7, Phase 2*). Currently, Accountability is weighted highest (30%) due to verification failures causing acute harm. In the future, organizations will be able to adjust (e.g., "For our legal team, Integrity should be weighted higher").

**Q: How does PAICE integrate with our LMS / HRIS?**
A: API access for HRIS integration planned for 2026. For pilot phase, assessment links are distributed manually or via email.

**Q: What's the pricing model?**
A: The current Individual assessment will remain *free forever*. Cohort assessments are planned to be free on limited trial basis, and also for approved Academic use. Additional paid products for Individuals, Cohorts (Teams, Trainings), and Enterprises will be announced in 2026. See *Section 7, Phase 2* of this whitepaper for more.

## About Validation & Evidence

**Q: Is PAICE scientifically validated?**
A: Partially. Individual-level measurement is operational and reliable. Framework face validity is strong (aligns with ISO/IEC 42001, NIST AI RMF). Organizational predictive validity (do scores predict work outcomes?) is being validated through pilot partnerships. Psychometric properties studies planned for 2026. We're transparent about being in Research Preview, see What "Research Preview" Means.

**Q: What evidence do you have that PAICE scores predict job performance?**
A: Not yet established. This is a key validation question for pilot partnerships. Early framework face validity is strong (aligns with ISO/IEC 42001, NIST AI RMF), but outcome correlation is TBD.

**Q: Can I use PAICE scores for hiring decisions?**
A: Not recommended yet. Predictive validity for hiring is unvalidated. If you're interested in pilot testing this (Track 3), contact Pilot [at] PAICE.work. Do not make adverse hiring decisions based on PAICE scores without proper validation and legal review.

**Q: Will PAICE be accepted by auditors for compliance evidence?**
A: Maybe. PAICE provides behavioral competence measurement aligned to ISO/IEC 42001 and NIST AI RMF, which is more defensible than training completion logs. However, regulatory acceptance depends on your industry, jurisdiction, and auditor. We're working toward this but can't guarantee it yet.

## About the Company

**Q: Why is PAICE a Public Benefit Corporation (PBC)?**
A: To ensure long-term mission alignment. PBC structure legally requires balancing profit with public benefit ("to enable safer and more effective People+AI collaboration by providing independent capability measurement."). This prevents pressure to prioritize growth over impact.

**Q: Who funds PAICE?**
A: Currently bootstrapped / pre-seed stage. Seeking strategic investors aligned with responsible AI and long-term infrastructure plays. No external funding as of November 2025.

**Q: Are you trying to replace human judgment with AI?**
A: No. PAICE measures human *collaboration* with AI, not replacement. The goal is humans using AI well, not AI replacing humans.

**Q: What's "Research Preview" mean for me?**
A: The assessment works and provides valuable insights, but formal organizational validation is in progress. We're transparent about what's proven vs. what we're establishing through pilot partnerships. Use PAICE for development and capability awareness, not high-stakes decisions (yet). See Research Preview Explanation for details.

# Appendix E: Limitations & Transparency

We believe transparent acknowledgment of limitations builds trust. Here's what PAICE cannot do, and what constraints remain.

## Current Constraints

**1. Language: English only**

- Multilingual support (Spanish, French, Portuguese, German) planned for 2026
- Non-English speakers currently excluded
- Cultural assumptions may favor English-speaking contexts despite attempts at neutrality

**2. Domain scope: Knowledge work focus**

- Early scenarios focus on business/knowledge work (reports, analysis, emails)
- Healthcare, legal, finance, creative verticals expanding in 2026
- Manual labor, physical tasks, purely creative work without AI tools = not applicable

**3. Model dependence: Claude-based**

- Currently relies on Anthropic's Claude for conversational assessment
- Single-vendor risk (model updates can shift scoring distributions)
- Multi-model ensemble planned to reduce dependency (Phase 2, 2026)
- Requires ongoing recalibration as LLMs evolve

**4. Validation stage: Research Preview 2025.11**

- Individual measurement is operational and reliable
- Organizational predictive validity is unvalidated (that's what pilots are for)
- Psychometric properties (test-retest reliability, etc.) are in progress
- Claims about organizational impact are projections, not proven outcomes
- Formal peer review in progress, not yet published

**5. Accessibility: Conversational format assumptions**

- Assumes literacy and digital fluency
- Extended time accommodations available but not fully tested across neurodiverse populations
- Screen reader compatibility untested (planned for 2026)

## What PAICE Cannot Do

❌ **Predict job performance comprehensively**

PAICE measures AI collaboration readiness, not overall job competence. A high PAICE score doesn't mean someone is a great employee—just that they're good at People + AI collaboration. Low scores don't mean someone is bad at their job—just that they need AI skill development.

❌ **Replace training**
PAICE identifies gaps but doesn't teach skills. It's diagnostic, not instructional. Organizations still need training programs, mentorship, and practice opportunities.

❌ **Work for all roles**
If your role doesn't involve AI tools (manual labor, creative work without AI assistance), PAICE is irrelevant. Don't force-fit measurement where it doesn't apply.

❌ **Guarantee compliance**
PAICE provides capability measurement aligned to ISO/IEC 42001 and NIST AI RMF, supporting compliance efforts. It does *not* replace legal review, regulatory consultation, or comprehensive risk management programs. Auditors may or may not accept PAICE scores as evidence—depends on industry and jurisdiction.

❌ **Eliminate all AI-related risk**
Even Exceptional-tier users can make mistakes. PAICE reduces risk by identifying capability gaps, but doesn't eliminate human error, malicious use, or AI system failures.

❌ **Provide real-time monitoring of AI use**
PAICE is a periodic assessment, not continuous surveillance. It measures capability at a point in time, not ongoing behavior. If you need real-time AI usage monitoring (e.g., DLP, policy enforcement), PAICE is not that tool.

## Known Limitations Under Active Investigation

**1. Score stability over time**

- Do PAICE scores remain stable if someone doesn't use AI for 6 months? Unknown.
- Test-retest reliability studies planned for 2026.

**2. Cross-cultural validity**

- Framework designed to avoid culture-specific knowledge, but untested at scale
- Bias audits planned across demographic groups (when self-reported)

**3. Construct validity**

- Do PAICE dimensions map cleanly to actual collaboration behaviors? Face validity is strong, empirical validation in progress.
- Correlation studies with related constructs (e.g., critical thinking, digital literacy) planned.

**4. Gaming resistance**

- Current anti-gaming measures (randomized scenarios, confidence checks, behavioral pattern analysis) are effective in early testing
- Sophisticated gaming (e.g., using AI to take the assessment) may not be fully detectable yet

**5. Tier boundary calibration**

- Current thresholds (Constrained 0-29, Proficient 50-69, etc.) are based on Research Preview data
- Pilot partnerships will refine boundaries based on outcome correlation

## Misuse Potential

**Risk: Scores used as employee ranking**
If organizations rank employees by PAICE score for promotion/comp decisions, this discourages learning (low scorers hide rather than improve). Mitigation: Non-punitive guidelines, privacy protections, emphasis on development not evaluation.

**Risk: Discrimination in hiring**
If PAICE scores are used to reject candidates without proper validation, this could introduce unfair bias. Mitigation: Explicit warning against adverse hiring decisions without legal review. Predictive validity for hiring is unvalidated—don't use it yet.

**Risk: False confidence**
High PAICE scores don't guarantee zero incidents. Organizations may over-rely on scores and under-invest in other risk controls. Mitigation: Transparent about what PAICE can/can't do, emphasize it's one tool among many.

**Risk: Teaching to the test**
If training programs optimize for PAICE scores rather than actual capability, measurement loses validity. Mitigation: Adaptive difficulty, randomized scenarios, behavioral observation (not just answer correctness).

## Ongoing Validation & Transparency Commitments

**Annual calibration studies**

- Re-validate scoring against reference scenarios
- Adjust tier boundaries if outcome data suggests miscalibration

**Quarterly bias audits**

- Analyze score distributions across demographic groups (when self-reported)
- Investigate disparate impact
- Recalibrate if bias detected

**Peer review**

- External academics and industry experts review framework
- Psychometric validation studies submitted to peer-reviewed journals
- Open to criticism and refinement

**Public transparency reports**

- Annual publication of validation findings, limitations, and score distributions

- No hiding negative results, if PAICE doesn't predict outcomes as hoped, we'll say so

## When NOT to Use PAICE

Don't use PAICE if:

- You need immediate, real-time AI monitoring (not a live surveillance tool)
- Your workforce doesn't use AI tools (no need for measurement)
- You expect perfect predictive validity (we're in Research Preview, validating now)
- You want a guaranteed compliance checkbox (regulatory acceptance is TBD)
- You need proven ROI before piloting (that's what pilots are for)

Do use PAICE if:

- You're deploying AI and need baseline capability measurement
- You want evidence-based training targeting (not generic AI courses)
- You need competence evidence for risk/compliance functions
- You're willing to co-validate and shape the tool through partnership
- You believe measuring People + AI collaboration is infrastructure worth building

---

# End