The Cost of Invisible AI Risk

A Board-Level Business Case for Measuring AI-Collaboration Reliability

Executive Brief

Sam Rogers, Founder of PAICE.work PBC

June 2026 • Version 1.0

For boards, CFOs, chief risk officers, and the executive sponsor of AI governance

Executive Summary

Your organization has spent on AI: licenses, training, and tools. It can report how many people are trained, how many seats are active, and how often the tools are used. It cannot report the one thing that determines whether that spending is safe: when the AI is wrong, do your people catch it? That blind spot is not a reporting gap. It is an unpriced liability, and it sits closest to the work that carries the most exposure: the judgment of licensed professionals in regulated functions.

This brief makes the financial case for closing it. The argument is simple. Liability for AI-assisted work lands on the organization, not the model. The cost of a single uncaught AI error reaching a client, a court, or a regulator is already being demonstrated in real rulings. And the rate at which errors pass unchecked is currently unmeasured, which means the exposure cannot be priced, managed, or reported to the board. **PAICE (People + AI Collaboration Effectiveness)** supplies the missing term: a measured, behavioral signal of whether your people actually catch AI error, scored like a credit score and tracked over time. It converts an invisible liability into a managed one.

The Exposure You Cannot Currently See

Every dashboard an executive sees about AI measures activity: adoption, usage, hours saved, tickets closed. None of it measures reliability. Activity tells you the tools are being used; it does not tell you whether the people using them would notice when the output is confidently wrong. The two are not correlated. A heavy user who trusts the model uncritically is a larger exposure than a light user who verifies everything.

PAICE's own data names the size of the blind spot. Across assessments, there is a consistent gap between how reliably people *believe* they verify AI output and how reliably they *actually* do under test, on the order of thirty points on a hundred-point scale. That gap is not a curiosity. It is precisely the region where organizational AI risk concentrates, because it is composed of people who are confident and wrong at the same time, and confidence is what determines whether a flawed AI output gets challenged or shipped.

> *You can measure system uptime. You cannot yet measure AI decision reliability. The gap between how reliably people believe they verify AI and how reliably they do is where organizational risk lives.*

Where the Exposure Becomes a Number

The cost of an uncaught AI error is no longer hypothetical, and the early rulings establish the principle that matters to a board: the liability attaches to the organization that deployed the work, not to the AI.

In **Mata v. Avianca** (U.S. District Court, Southern District of New York, June 2023), lawyers submitted a brief citing six judicial decisions that did not exist; ChatGPT had fabricated them, and the attorneys had not verified them. The court sanctioned the lawyers and their firm under Rule 11. In **Moffatt v. Air Canada** (British Columbia Civil Resolution Tribunal, February 2024), the airline was held liable for wrong information its own chatbot gave a customer; the tribunal rejected as a “remarkable submission” the argument that the chatbot was a separate entity responsible for its own actions. The organization owns what its AI-assisted work product says.

The direct awards in these early cases were small: a five-thousand-dollar sanction in Mata, an eight-hundred-dollar refund in Moffatt. That is the trap. The headline number understates the real cost by an order of magnitude, because it excludes the legal fees, the remediation, the reputational damage, and the precedent each ruling sets for the next. And these are not isolated incidents. Appellate courts and bar regulators have since acted on the identical failure: a discipline referral by the Second Circuit in Park v. Kim (2024), Florida Bar disciplinary proceedings in the matter of Neusom (2024), and fresh sanctions in Wadsworth v. Walmart (District of Wyoming, 2025). For a licensed professional, the consequence escalates past a fee sanction to a disciplinary finding and a license at risk. The cost scales with the stakes of the work, and the work that uses AI most is increasingly the work with the highest stakes.

> *The liability lands on the organization, not the model. The small awards in the first cases understate the true cost, because the precedent does not stay small.*

Why Current Spending Does Not Touch It

The reasonable executive response is: we already invest in this. We run AI-use training, we hold licenses with reputable providers, we have an acceptable-use policy. None of that addresses the exposure, because all of it measures inputs. Training completion records attendance, not competence. License counts record access, not judgment. Usage dashboards record activity, not reliability. A workforce can be fully trained, fully licensed, and fully active while remaining unable to catch the errors that create liability. The spend is real; it simply does not reach the behavior that carries the risk.

This is the same pattern that financial institutions learned about operational risk and that security organizations learned about human factors: you cannot manage what you only measure by proxy. The proxy here is usage; the risk is behavioral; and the two have been allowed to stand in for each other because the behavioral measure did not exist.

Pricing the Exposure

A board does not need a precise loss figure to act; it needs to know the exposure is real, directional, and currently unmanaged. The exposure can be framed simply, and the framing shows exactly which term is missing.

Term	What it is	Observable today?
Population at risk	People doing AI-assisted work in roles that carry liability or regulatory duty.	Yes: HR and usage data.
Verification-failure rate	The share of AI errors that pass the human reviewer unchallenged.	No: this is the blind spot.
Cost per uncaught error	Liability, remediation, and reputational cost when one error reaches a client, court, or regulator.	Partly, and only after the fact.
Exposure	Population × verification-failure rate × cost per uncaught error.	Not computable without the middle term.

Table 1: The AI-risk exposure framework. The organization can already estimate the first and third terms; the verification-failure rate is the one it cannot see, which is why the exposure has gone unpriced.

PAICE supplies the missing middle term. By observing whether people actually catch deliberately injected AI errors, it produces a measured verification-reliability signal per role and per cohort, the leading indicator of the failure rate. With it, the exposure stops being unknowable and becomes a number that can be tracked, targeted, and reported, the same way a credit score makes default risk legible without predicting any single default.

The Instrument and the Decision

PAICE is a behavioral assessment of how people work with AI. It scores observed conduct, not self-report, on a 0–1000 scale, with the heaviest weight on whether the person catches AI error and takes ownership of AI-informed decisions. It does this without storing conversation content or individual identities, so it produces the signal a board needs without creating a surveillance liability or a data-protection one. The output is a **credit-score-style measure of AI-collaboration reliability**: legible to a non-technical decision-maker, comparable across the organization, and trackable over time.

The decision in front of the board is not whether to spend more on AI. It is whether to make the largest unpriced AI risk visible before an incident prices it for you. Regulators are moving the same direction: evidence of human oversight is becoming an obligation, not a courtesy, which means the behavioral signal will soon be something the organization is asked to produce

regardless. The choice is to build that visibility deliberately and ahead of need, or to acquire it reactively after the first uncaught error becomes the organization's own case study.

The next step

Run a baseline assessment of one high-stakes cohort. It produces a verification-reliability signal for that group in weeks, at a cost far below a single uncaught-error incident, and turns the exposure framework above into a real number the board can act on.

To scope a baseline, contact **info@paice.work** or visit paice.work.

This document is provided for informational purposes only and does not constitute legal, financial, or compliance advice. Case references are drawn from the AI Incident Law database (aiincidentlaw.org): AIEL-2023-002 (Mata v. Avianca, S.D.N.Y., 22 June 2023), AIEL-2024-001 (Moffatt v. Air Canada, 2024 BCCRT 149), AIEL-2024-003 (Park v. Kim, 2d Cir. 2024), AIEL-2024-008 (In re Neusom / Florida Bar, 2024), and AIEL-2025-004 (Wadsworth v. Walmart, D. Wyo. 2025). Descriptions are summaries, not legal analysis. The exposure framework is illustrative and not a loss model.

Appendix: Litigation and Incident Patterns from the AI Incident Law Corpus

The five matters cited in the body are not isolated examples. They sit inside a curated corpus of public AI-incident matters maintained at aiincidentlaw.org and used as the source of record for this brief. As of June 2026 the included set contains fifty matters spanning U.S. federal trial and appellate courts, multiple state appellate courts, a Canadian tribunal, and federal regulators (FTC, EEOC). The patterns below are what the corpus actually shows, not a projection.

Incident pattern	Share of corpus	Representative matter
Fabricated authority in court filings (citations, quotations, or holdings invented by a generative model and submitted without verification)	37 of 50	Mata v. Avianca (S.D.N.Y. 2023); Park v. Kim (2d Cir. 2024); Wadsworth v. Walmart (D. Wyo. 2025)
Deployer liability for chatbot output to a customer or counterparty	1 of 50	Moffatt v. Air Canada, 2024 BCCRT 149
Facial-recognition false match leading to wrongful arrest or detention	4 of 50	Williams v. Detroit (E.D. Mich.); Parks, Woodruff, Murphy matters
Algorithmic eligibility, benefits, or allocation harm at scale	3 of 50	MiDAS Michigan unemployment system; Arkansas RUGs care-allocation; SafeRent tenant scoring
Hiring and screening discrimination by automated tools	3 of 50	EEOC v. iTutorGroup; Workday HiredScore collective action; IBM rehire screening
Regulator-led enforcement against an AI deployment	2 of 50	FTC In re Rite Aid (five-year facial-recognition ban); EEOC v. iTutorGroup ($365,000)
Output and training IP claim against a generative AI provider	1 of 50	CNN v. Perplexity

Table A1: Incident-pattern distribution across the fifty included matters in the AI Incident Law corpus as of June 2026.

The dominant pattern is fabricated authority in court filings: roughly three of every four included matters. The shape is consistent. Counsel uses a generative model (most often ChatGPT, but increasingly a vendor-branded legal research product such as CoCounsel, Lexis+, or Fastcase) to draft a brief, does not verify the citations, and files. Opposing counsel or the court catches the fabrication. The remedy escalates with recurrence: a first-time matter may end at a Rule 11 fee

sanction in the low thousands; repeated or aggravated conduct draws fee-shifting awards into the tens of thousands, public reprimands, bar referrals, pro hac vice revocation, and, in at least one matter, a court-ordered series of bar-association speaking engagements. Thirty-four of the fifty included matters end in a formal sanction.

The Moffatt principle (the deployer owns what its AI says) generalizes beyond legal services. The corpus shows the same liability route through customer-facing chatbots, eligibility systems run by state agencies, hiring tools licensed from vendors, and surveillance systems operated by retailers. The organization that puts the AI in front of the decision is the organization that answers for the error, regardless of where the model was built or whether a human reviewed the output.

Tool attribution is drifting. Early matters named ChatGPT explicitly. More recent matters increasingly involve unnamed generative AI or brand-name legal-research products. Courts have begun rejecting attorney attributions to specific vendors when the fabrication pattern does not match that vendor's behavior. The implication for risk owners is that the question is not which tool was used; it is whether the person using it caught the error.

Matter	Monetary or remedial outcome	Significance
Mata v. Avianca (S.D.N.Y. 2023)	$5,000 Rule 11 sanction	First widely-reported fabricated-authority sanction
Moffatt v. Air Canada (BCCRT 2024)	~$812 consumer award	Deployer-liability principle established
EEOC v. iTutorGroup (E.D.N.Y.)	$365,000 plus injunctive relief	Regulator action on algorithmic hiring
MiDAS (Michigan)	$20 million settlement	Mass-harm aggregate from automated decisions
In re Rite Aid (FTC 2023)	Five-year ban on facial recognition	Regulator removal of an AI capability
Wadsworth v. Walmart (D. Wyo. 2025)	$7,000 fee sanction plus ten mandatory bar speaking engagements	Behavioral remediation added to monetary penalty
Brigandi/Murphy fee shift	$94,704 attorney-fee award plus claim dismissal	Largest single fee shift in the corpus

Table A2: Representative monetary and remedial outcomes from the corpus. Outcomes range from low-three-figure consumer awards to mid-six-figure fee shifts, multi-million-dollar mass-harm settlements, and outright capability bans. The cost is not bounded by the size of the first sanction.

Across all fifty matters the through-line is the one the body argues: liability lands on the organization that deployed the work product, the cost scales with the stakes of the work, and the determinative behavior is whether the human in the loop caught the error before it left the building. The corpus is the receipts.