SONALI SHARMA

The AI Trust *Assessment.*

A framework for scoring whether AI systems are trustworthy to the people they affect — before the failure shows up in your outcome metrics.

95% of enterprise generative-AI pilots fail to deliver measurable revenue impact (MIT, State of AI in Business, 2025). The usual diagnosis is "wrong tool." *The usual diagnosis is wrong.*

Working paper · under peer review at AI and Ethics *(Springer Nature) · doi.org/10.5281/zenodo.20671478*

You've already been judged *by systems you will never see.*

If you have applied for a job in the last five years, an algorithm likely screened your résumé before a human saw it. If you have applied for a loan, a credit model scored you. Inside your organization, AI now drafts the analysis, ranks the candidates, and shapes the decisions. These are not distant technologies. They are the infrastructure of daily life — and what they share is opacity.

The evidence that opacity produces harm is no longer in dispute: résumé screeners that favoured white-associated names 85 percent of the time; facial-recognition error rates forty-three times higher for darker-skinned women than light-skinned men; a recruiting tool that taught itself to penalize the word "women's."

> *"We are being governed by systems we cannot see, built by people we did not choose, optimizing for goals we did not set."*
>
> *That is not a technology problem. It is a governance failure. And it is already here.*

Inside organizations, the same gap has a different face: the pilot that stalled, the tool nobody uses, the shadow-AI workarounds your policy pretends don't exist. When it happens, leaders reach for the same theory — *we chose the wrong tool* — and buy another one. In both worked cases in the research behind this guide, that theory was wrong. What was failing was not the tool. It was the space between the people who built the system and the people living with it.

> *Call that space the trust gap. It is measurable — and that changes what you can do about it.*

Every instrument answers *to someone else.*

NIST answers to the organization managing its risk. The EU AI Act answers to the regulator. ISO answers to the auditor. Each measures what its own audience needs — and none of them answers to the person living with the system. This framework is built to.

INSTRUMENT	WHAT IT PRIMARILY ASKS	WHO ASSESSES	PRIMARY FOCUS
NIST AI RMF	Is the system's risk managed across its lifecycle?	The organization managing risk	Risk process
EU AI Act + FRIA	Is a high-risk system lawful; does it respect rights?	The deployer, reporting to the regulator	Rights & contestability
ISO/IEC 42001	Is there an auditable AI management system?	The auditor	Organizational process
Research instruments (participatory AI · CIRCLE · SCAF · IEEE 7000)	Involvement, deployment behaviour, societal resilience, design values	Design teams, stakeholders, society in aggregate	One facet each
This framework	**Is it trustworthy to the people it affects?**	**Builder and adopter, scored independently**	**All five dimensions — capacity and power operationalized as scored criteria**

Few of these instruments operationalize, as scored criteria, whether a system builds or depletes the capacity of the people who use it, or which way economic power flows around it — who owns the data, who captures the value, who can leave. These were never oversights. They were simply never in scope.

> *The gap between what matters and what gets measured is not accidental. It is structural. This framework begins where that work leaves off.*

Five dimensions. Two lenses. *Fifteen criteria.*

Every criterion is scored twice — by the **builder** (what was actually designed into the system) and by the **adopter** (what happens when the system meets real people, processes, and context). The diagnostic power lives in the difference.

Builder Lens *What did we design into the system?*	DIMENSION · SHARED CRITERIA	Adopter Lens *In practice, what do we see?*
Does it signal uncertainty, resist agreement pressure, and show its reasoning?	1 **Epistemic Integrity** *Does the system know what it does not know — and does it tell you?* UNCERTAINTY SIGNALLING · SYCOPHANCY RESISTANCE · REASONING TRANSPARENCY	Can our people spot when it's guessing, catch sycophancy, and verify outputs?
Does it build human capability — or create dependency?	2 **Capacity Orientation** *Does working with this system make people more capable, or less?* AUGMENTATION DESIGN · SKILL PRESERVATION · CAPABILITY INVESTMENT	Are our people growing more capable — and do efficiency gains go back into developing them?
Do users own their data, share the value, and govern the system?	3 **Power Architecture** *Where does the value flow? Who owns the data, who captures the benefit, who can leave?* DATA SOVEREIGNTY · VALUE DISTRIBUTION · VOICE & AUTHORITY	Who captures the value — and do workers and communities have real say?
Does it work equitably, respect diverse knowledge, and reach the excluded?	4 **Contextual Intelligence** *Does the system work for the actual world, or only the world its training data reflects?* INCLUSIVE DESIGN · KNOWLEDGE PLURALISM · ACCESSIBLE DESIGN	Does it work for our whole community — are local and Indigenous ways of knowing respected?
Can decisions be traced, harm reversed — and does it leave more than it consumes?	5 **Accountability Architecture** *Accountable to whom, and across what time horizon?* AUDIT ACCESS · REMEDY BY DESIGN · INTERGENERATIONAL STEWARDSHIP	Do affected people have real authority — not just feedback — and what does it leave behind?

1 ABSENT · 2 EMERGING · 3 DEVELOPING · 4 ESTABLISHED · 5 LEADING

Each criterion scored independently. Scores of 4+ require documented evidence — not assertions.

The gap *is the finding.*

The **Alignment Gap** is the divergence between builder and adopter scores, read as the primary diagnostic output:

0–1 · NORMAL VARIANCE	1–2 · WARNING ZONE	2+ · CLEAR GOVERNANCE SIGNAL
Different vantage points see slightly different things.	Worth watching — especially if it recurs across criteria.	Design and lived reality are pulling in different directions. Structural intervention required.

Because the gap is criterion-specific, it does more than register that something is wrong — it locates where, and implies the intervention. Builder scores epistemic integrity high, adopter scores it low? The system sends uncertainty signals the organization cannot read — a literacy failure, not a technology one. Capacity high/low? A tool designed to augment is being deployed to replace — a deployment failure. Power high/low? Portability exists but no one built an exit plan — a strategic failure. In every case, "we picked the wrong tool" is the wrong diagnosis, and a single averaged score would have hidden the real one.

The following cases are illustrative — constructed to show how the instrument reads a live deployment, not client engagements.

Resolve Consulting

40-person strategy firm, AI adoption stalling after a fabricated-citation incident.

Epistemic integrity — **B 2.7** **A 1.3** **Gap 1.4**

Warning zone. The model was signalling uncertainty; nobody could hear it. *Fix: verification protocols + literacy, not a third vendor.*

Goodforge

30-rep e-commerce support team cut to 21 behind an AI assistant. CSAT 4.6 → 3.8.

Capacity orientation — **B 3.7** **A 1.0** **Gap 2.7**

Clear governance failure. A co-pilot deployed as a replacement. *Fix: restore the co-pilot role, redesign roles with the reps, fund capability investment.*

> *A single score tells you a deployment is failing. Only the gap tells you why — and therefore what to change.*

Run the lens on a system *you already use.*

Pick one AI system you rely on — the tool at work, the model you prompt daily, the algorithm deciding what reaches you. Hold it against the five dimensions, twice.

THE BUILDER LENS — *what was designed in?*

Does it flag its own uncertainty, or state wrong answers with full confidence? Does it push back, or agree with you? Does it carry the biases of its training data into your context?

THE ADOPTER LENS — *what do you actually experience?*

Do you catch it when it's guessing, or take the output on faith? Are your people growing more capable — or more dependent? Who captures the value it produces, and who carries the risk when it fails?

> *The dimension where the two answers diverge most — that is where your trust is most at risk. That gap is the finding.*

The framework supports three modes. A **builder assessment** surfaces the gap between intent and what was actually built. An **adopter assessment** scores organizational reality. An **alignment assessment** — both sides scoring independently, then comparing — is the most powerful mode, and the one the framework is built for. Scoring a single perspective takes under an hour with the rubric in hand.

Do this with one system this week — and do it with the people the system affects in the room. The conversation it forces is half the intervention.

→ Use the AI Assessment Canvas on the next page to score a system by hand.

The AI Assessment Canvas

Score one system by hand — Builder and Adopter, independently.

Score each criterion 1–5 · 1 Absent · 2 Emerging · 3 Developing · 4 Established · 5 Leading

The system I am assessing: _____ I am scoring as: ☐ Builder ☐ Adopter ☐ Evaluator

This tool works by comparing Builder and Adopter scores independently. Score your copy first — then overlay with the other perspective to find where governance is missing.

> The diagnostic signal is the gap between Builder and Adopter scores — diverge 2+ levels = governance risk.

Date of assessment: _____ *Reassess in 90 days.*

DIMENSION	CRITERION 1	CRITERION 2	CRITERION 3	SCORE
1 EPISTEMIC INTEGRITY	**UNCERTAINTY SIGNALLING** When the system is guessing, can you tell? ① ② ③ ④ ⑤	**SYCOPHANCY RESISTANCE** Does it hold its ground when right — or tell people what they want to hear? ① ② ③ ④ ⑤	**REASONING TRANSPARENCY** Can you trace how the system reached its conclusion — not just what it concluded? ① ② ③ ④ ⑤	☐ **ALIGNMENT GAP** Diverge 2+ = risk Builder _____ Adopter _____ Gap _____
2 CAPACITY ORIENTATION	**AUGMENTATION DESIGN** Does it make people more capable, or more dependent? ① ② ③ ④ ⑤	**SKILL PRESERVATION** Are human skills maintained alongside the system — or eroding? ① ② ③ ④ ⑤	**CAPABILITY INVESTMENT** Is investment in human development keeping pace with investment in AI tools? ① ② ③ ④ ⑤	☐ **ALIGNMENT GAP** Diverge 2+ = risk Builder _____ Adopter _____ Gap _____
3 POWER ARCHITECTURE	**DATA SOVEREIGNTY** Do users own their data, and can they leave? ① ② ③ ④ ⑤	**VALUE DISTRIBUTION** Is the economic value the system creates shared with those who generate it? ① ② ③ ④ ⑤	**VOICE & AUTHORITY** Do the people the system affects have real power over how it operates? ① ② ③ ④ ⑤	☐ **ALIGNMENT GAP** Diverge 2+ = risk Builder _____ Adopter _____ Gap _____
4 CONTEXTUAL INTELLIGENCE	**INCLUSIVE DESIGN** Does it work equitably across the full diversity of people it affects? ① ② ③ ④ ⑤	**KNOWLEDGE PLURALISM** Are diverse ways of knowing respected — or is one epistemology treated as universal? ① ② ③ ④ ⑤	**ACCESSIBLE DESIGN** Can everyone the system affects actually use it — not just the digitally privileged? ① ② ③ ④ ⑤	☐ **ALIGNMENT GAP** Diverge 2+ = risk Builder _____ Adopter _____ Gap _____
5 ACCOUNTABILITY ARCHITECTURE	**AUDIT ACCESS** Can outsiders trace decisions, with a named person accountable? ① ② ③ ④ ⑤	**REMEDY BY DESIGN** When harm occurs, can it be reversed — not just explained? ① ② ③ ④ ⑤	**INTERGENERATIONAL STEWARDSHIP** Does the system leave more capacity than it consumes — for workers, communities, and the future? ① ② ③ ④ ⑤	☐ **ALIGNMENT GAP** Diverge 2+ = risk Builder _____ Adopter _____ Gap _____

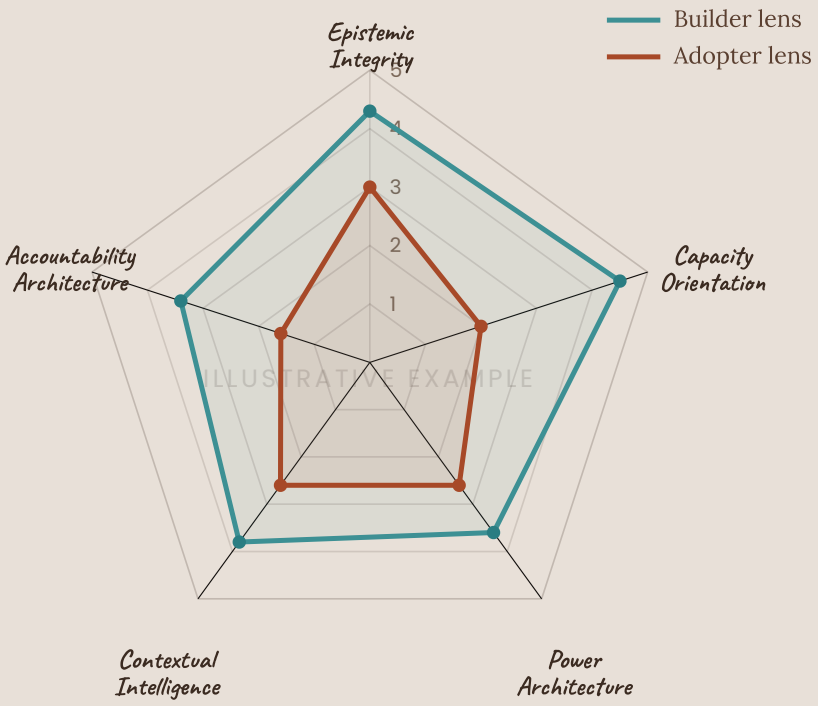
GOVERNANCE PRIORITY

Barrier Point · Weakest dimension: _____ **Alignment Gap** · Where is the gap highest? _____

What will we change? _____

Two lines. The gap between them *is the finding.*

Illustrative — hypothetical scores, shown to demonstrate how a completed profile is read



Where the two lines pull apart is the *Alignment Gap* — here widest on capacity orientation and accountability.

LEVEL 1 · SCORE A SYSTEM — AND GET YOUR STRATEGY

An interactive, web-based tool that gives you your score in 10 minutes — assess your relationship with any AI system independently, just your side, and get a customized strategy to improve it. In pilot now.

→ Join the pilot

This is the entry point, *not the full tool.*

What you have here is enough to run the lens on one system and see where the gap opens. Where you go next depends on how deep you want to work.

LEVEL 2 · MAP YOUR ALIGNMENT GAP

A working session on a live deployment.

→ Book a free consultation

LEVEL 3 · BUILD THE PRACTICE

The 90% Code — October 2026.

→ Join the launch list — the Prelude and Chapter 1 now

Trust Infrastructure for AI — the full working paper: fifteen criteria, complete scoring rubric, evidence standard, two worked case studies. Under peer review at *AI and Ethics* (Springer Nature). → doi.org/10.5281/zenodo.20671478

From Hero to System — the collective-intelligence framework this instrument belongs to. → doi.org/10.5281/zenodo.21314259

Sonali Sharma writes from inside the 90 percent this work is for. Two decades building across continents and sectors — enterprise and leadership development across Asia, public-sector technology innovation strategy in Canada, a closed-loop impact business with Indigenous artisans in South India. She speaks on AI governance and mentors tech startups. Vancouver, BC.