Closing the AI Trust Gap

By Alan Hilburg, Said Saillant and Patricia Caporaso
September 2024

Not just bias and hallucinations—but a failure to question. Why curiosity, scrutiny, and accountability are AI's last best hope.

The world's most advanced algorithms can diagnose cancers, draft contracts, and steer cars—yet millions hesitate to rely on them. That hesitation is the AI trust gap: the distance between what artificial intelligence *can* do and what people are willing to let it do. Closing that gap hinges on a single habit of mind: validation-driven curiosity—the discipline to question, verify and understand before acting.

Last December's Global Partnership on AI (GPAI) Summit in Belgrade offered a quiet but clarifying lesson. Although the agenda stressed international cooperation, only a handful of audience questions made it onto the stage—a reminder that, without open dialogue, trust cannot take root. The real gap on display was not a shortage of answers, but of opportunities to ask.

If we want AI's transformative promise to reach healthcare, finance, and climate action, we must replace that silence with relentless inquiry and shared accountability—asking the hard questions that earn AI the credibility it needs to thrive.

Both real and perceived risks fuel the AI trust gap. Biased hiring tools, chatbots that invent legal citations, and self-driving cars that still miss pedestrians remind the public that AI can be wrong—and dangerous. No wonder a recent global survey found 61% of adults report low or moderate trust in AI. In machine-learning jargon, *hallucination* describes plausible yet false outputs, simulating reality without grounding in fact. When an AI "hallucinates" a medical diagnosis or a news headline, the consequences are real. Facing dangers that range from bad loans to botched diagnoses, we need more than better code—we need a shared plan for trust. Building that plan requires everyone at the table.

No single actor can close this gap alone. Developers, regulators, and users must work in concert to build an ecosystem of trust—one that unlocks AI's potential without compounding its risks.

Developers hold the keys to creating systems that promote transparency and informed use. By embedding tools that encourage users to question AI outputs, they can transform opaque technologies into trusted, collaborative partners.

Regulators must legislate strong ethical standards that champion fairness, security, and accountability. Clear guidelines and effective oversight help align AI with societal values rather than short-term gains.

Users must have the power to interrogate AI—probing its outputs, verifying claims, and embracing its imperfections. By treating AI as a partner in thinking rather than an infallible oracle, users actively guide its evolution.

So, what values should steer that ecosystem? Start with four pillars—each activate by curiosity:

- **Accountability**: Clear lines of responsibility and meaningful recourse when things go wrong.
- **Integrity**: Ethical design that resists hidden agendas or data manipulation.
- **Reliability**: Consistent, peer-tested performance across contexts.
- **Vulnerability**: Open admission of limits and a willingness to say, "I don't know."

By relentlessly asking *Who answers for errors?* or *What edge cases break this model?* we keep our reliance on AI informed, not blind—much like checking street signs even when GPS shows the way. "Trust, but verify" resonates in the AI age—placing faith in potential while demanding accountability.

Fine words—but they matter only if we can apply them. Enter **the AI Challenge Protocol**—three quick checks that developers can embed, regulators can codify, and citizens can practice:

1. **Engage.** Pause before accepting an output. *Ask who built the system, on what data, and for whose goals. Example: "Which hospital records trained this diagnostic model?"*

2. **Question.** Probe for blind spots or skewed priorities. *Example: "If I swap the applicant's zip code, does the loan rate change?"*
3. **Test.** Cross-check with trusted sources or independent evidence. *Example: "Do SEC filings confirm the revenue figure ChatGPT just gave me?"*

Practiced together, these steps turn passive consumers into active partners and remind us that trust is earned through verification.

Yet scrutiny cannot work in the dark. Verification depends on visibility—and that's where transparency comes in. Open-source code and public datasets are steps in the right direction, but people must also understand how the parts fit together. Tools that support meaningful inspection—like well-designed audit trails or interpretable summaries—can help. But exposing the gears of a watch means little if we cannot grasp how they tell time. Transparency must lead to understanding, not just disclosure.

That's where curiosity reenters—not as a soft virtue, but as a strategic necessity. It's curiosity that drives us to ask, "How does this system decide?" or "What might it be missing?" These questions aren't just philosophical—they're practical tools for ensuring accountability. They move us from passive acceptance to active engagement.

Developers can build systems that invite scrutiny through thoughtful design. Regulators can demand disclosures that clarify—rather than obscure—how AI systems are built, trained, and deployed. And users, guided by informed curiosity, can push AI to serve as an extension of human judgment, not a replacement for it.

Three tests will still need to answer:

1. **Who is accountable when AI systems fail?** Developers, policymakers, or users—and how is that accountability enforced?
2. **How will AI stay truthful and fair when profit or convenience says otherwise?** What oversight protects the public good?
3. **Can citizens shape AI governance?** Have existing frameworks sidelined their voices, and how can we embed genuine public participation?

Trust in AI isn't automatic; it must be earned—step by step through open inquiry, balanced debate, and relentless validation. That's how we close the AI trust gap: with curiosity-driven questions, honest conversations, and a shared commitment to excellence. Because the true measure of progress in AI won't be how fast we move, but how wisely we steer.

Curiosity is how we close it.