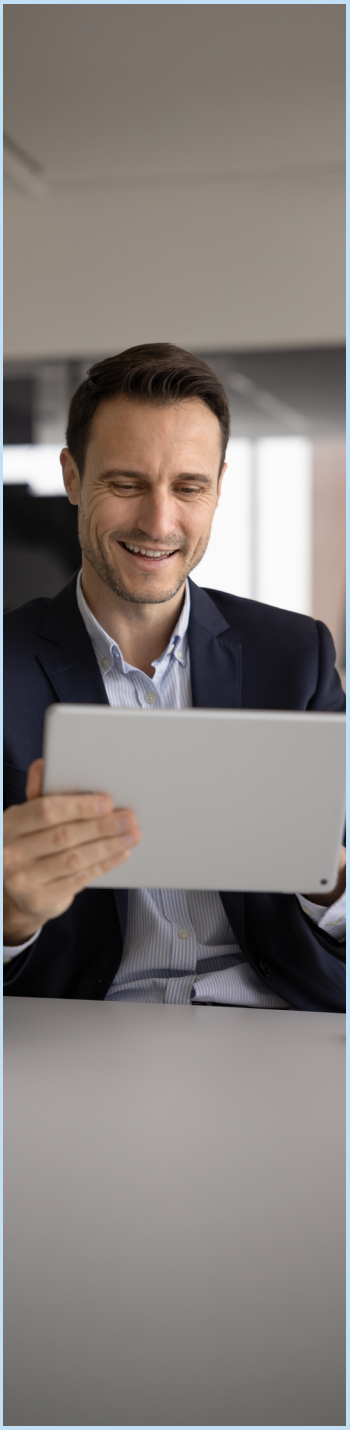
VIEW POINT



BUILDING TRUST IN AI THROUGH TRANSPARENCY AND HUMAN OVERSIGHT

Abstract

As AI adoption accelerates across industries, trust in its output remains fragile because transparency, accountability, and human oversight have not kept pace with deployment. This article argues that trustworthy AI cannot be achieved through model capability or regulatory compliance alone; it requires explainability that reaches decision-makers, verifiable data provenance, continuous governance contracts, and human checkpoints designed to identify errors, bias, drift, and vendor-risk gaps. By treating transparency and oversight as ongoing operational disciplines rather than one-time compliance exercises, organisations can build AI systems that people are more willing to understand, question, and responsibly rely on.

Infosys®
Navigate your next

Transparency: Implications and nuance

"Transparency" gets used loosely in AI discourse, often as a synonym for good intentions. Introducing model explainability increased how quickly senior leadership trusted and accepted AI outputs. There is evidence to show that using explainability techniques such as SHAP and LIME specifically can build internal trust. While regulatory requirements may not ask for it across all jurisdictions, showing leadership how a decision was reached could nudge them towards acting on it.

Another notable observation is that transparency between an AI provider and a client created the opening for the client to request additional human-in-the-loop checkpoints. Transparency, in that case, would be the precondition for oversight.

At the organisational level, data transparency and outcome transparency are related but distinct problems. Model transparency can coexist with data transparency failure. A system can explain its outputs (model transparency) while still drawing on data of unverifiable origin (data provenance failure) and producing outcomes that no one is tracking against real-world consequences (outcome transparency failure). Poorly documented training data whose real-world impact is unmonitored is not a transparent system in any meaningful sense of the word

Solving one problem does not solve the other.

The other dimension that governance frameworks often miss is attribution. Decision-makers need to know not just what the AI concluded, but what it drew on, what confidence level it holds, and what knowledge, verified by a human, informed its reasoning.

Retrieval-Augmented Generation (RAG), which entails combining a language model capability with a curated, human-verified knowledge repository, has proven to be successful in experiments done by some of the largest AI tool providers. The result was not that the model became more capable. It was that the people using it became willing to rely on it.

Trustworthiness as a continuous contract

Organisations think about AI trustworthiness as something declared at deployment and validated once, rather than something that must be actively maintained over time.

A more useful framework is to think of trustworthiness as a set of explicit contracts, containing defined commitments that an AI system must fulfil to continue functioning. Those commitments might include accuracy, fairness, technical robustness, and transparency, each specified to the degree required for the system's particular context. The contracts must be specific to what the system does and whom it affects.

When an AI system drifts from the commitments that define it, when its accuracy degrades, when its fairness metrics shift, or when its outputs no longer reflect the data conditions it was trained on, it is not the same system in any meaningful sense. An AI system that has changed in ways that compromise its trustworthiness profile has changed in ways that matter to everyone who relies on it.

Retraining, version updates, and architectural changes all create checkpoints where trustworthiness must be re-verified. A governance programme must account for this.

Human oversight in AI

The EU AI Act mandates human oversight for high-risk AI systems. What effective oversight looks like in practice has not been defined yet.

Stakeholders in technical roles have consistently called for human checkpoints at defined decision nodes. However, humans in the loop introduce their own biases too. If the person providing oversight is financially incentivised toward a particular outcome, their oversight is no longer neutral.

Effective human oversight in AI requires both the technical design of the checkpoint and the right incentive structure for the personnel staffing it. Outputs from LLMs and other AI applications are promising, fast, but prone to specific and recurring categories of error, and require active supervision.

Accountability under fragmentation

Modern AI systems are built on chains of vendors, sub-processors, and pretrained models from other organisations. When something goes wrong, the responsibility for that failure is difficult to locate.

Many practitioners working with enterprise AI systems report finding themselves in a position of having to trust third-party compliance assurances with no mechanism to verify them independently. A site visit to a vendor's data centre can confirm the aspects that are visible. One cannot confirm the vendor's internal practices, what the practices of their sub-processors are, or the characteristics of the pretrained model components that their offerings are based on. Regulatory frameworks allocate responsibility to the provider who builds or modifies the system, and independent audits are not feasible.

What does seem to reduce the damage, without resolving the underlying tension, is treating data provenance and audit trails as continuous operational infrastructure rather than point-in-time compliance exercises. Organisations that invest in automated data lineage tracking report better outcomes: fewer wasted hours, fewer misunderstood inputs, and cleaner institutional memory when staff change roles. While this does not eliminate the third-party trust issue, it does make the parts of the system an organisation controls more accountable.

What building trust actually requires

Transparency has to be specific. Explainability techniques that show how a decision was reached measurably increase the willingness of both leadership and end users to act on AI outputs, but only when that explanation reaches the people who need to trust the system, not just internal technical teams.

Oversight has to be designed against incentives. A human checkpoint staffed by someone with a financial stake in the outcome, or insufficient technical grounding to evaluate what they're reviewing, is not real oversight.

Attribution and provenance do more practical work than abstract trust-building language. Knowing where an output came from, what it drew on, and how confident the system is in it gives people something concrete to evaluate, rather than asking them to simply have faith.

Accountability for third-party and vendor risk remains unresolved at an industry level. The prevalent position is that they have built partial mitigations, such as audit trails, due diligence processes and contractual accountability clauses, but not a complete answer.

End note

Trust in AI can only be nurtured with time. It is built through deliberate transparency about how systems reach their outputs and through human oversight that is designed with enough rigour to actually catch what it's meant to catch. The gap between AI adoption and AI trust will not close on its own. Closing it requires the ongoing work of provenance tracking, explainability infrastructure, and oversight roles staffed and incentivised to do the job properly.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next