

AI Guardrails: Strengthening Trust and Safety for Sovereign AI

A visual guide for Indian banking leaders

RESPONSIBLE AI
STRONGER BANKS
A MORE INCLUSIVE INDIA
TRUST TODAY.
A STRONGER TOMORROW.

55

The question is not whether the model is accurate. It is what the system is allowed to do, what happens when it should not, and where the bank can prove either.

SAFER AI.
STRONGER CUSTOMERS.
BRIGHTER INDIA.



AI System

Model / Agent / Application



Guardrail

Rules / Monitoring / Control



Customer / Account /
Record

Real-world impact



Audit Trail

Logs / Evidence / Reporting

1 What an AI guardrail does

- Constrain** — Keeps the model or agent inside its approved scope
- Monitor** — Checks every output and action as it happens
- Record** — Creates an auditable trail, including blocked decisions



For regulated institutions, the evidence of the control matters as much as the control itself.

2 Where AI already operates in Indian banks

- Fraud detection & AML monitoring
- Credit assessment & cash-flow-based lending
- Customer service
- Early-warning systems for stressed accounts
- Document processing for underwriting & collections



The more decision a system owns, the more the control around it has to do.

4 What slows safe AI adoption?

- Fragmented customer, transaction, and risk data
- Legacy core systems with incomplete real-time context
- Supervision teams still building oversight for acting AI



The controls that hold are built into the layer the AI runs on.

3 Generative AI vs Agentic AI



Generative AI
GOVERN THE OUTPUT

Example: answering a loan-related query. Prevents invented rates, ineligible offers, or data exposure, and holds or escalates edge cases before a response is sent.



Agentic AI
GOVERN THE ACTION

Example: flagging a transaction or restricting a card. Requires a permission check before the action proceeds, with gating in real time, inside the system.



A guardrail built to review what a model says is not enough for a system that can act.

5 Why sovereign infrastructure matters

- Bank-controlled** — Guardrail, data, and decision record stay inside bank-controlled infrastructure
- Architecture, not roadmap** — Governance depends on the bank's architecture, not a vendor's roadmap
- DPDP-ready** — Supports purpose-bound data use under India's Digital Personal Data Protection Act
- Trust follows infrastructure** — Trust depends on where the guardrail actually runs

6 Putting guardrails into practice The Governance pillar of iStreet's Sanjeevani of AI™ framework

- 01** Define what the AI is allowed to do
- 02** Check each output and action in real time
- 03** Flag or escalate exceptions before impact
- 04** Keep control and evidence in the same record



For BFSI, trust in AI means trusting the guardrails around it — and knowing where they run.