RiskCounts

Responsible AI in the Indian Financial System

The Seven Sutras of FREE-AI

6 September 2025

Introduction: The Strategic Imperative of AI Governance

Artificial Intelligence has moved beyond experimental pilots to become a strategic lever across Indian financial services. Banks, NBFCs, insurers, and fintechs increasingly rely on Machine Learning and Artificial Intelligence to support credit, risk management, fraud detection, underwriting, payments, customer engagement, regulatory compliance and other internal processes. With this influence comes great responsibility. Errors, bias, opacity, or unintended consequences in AI-driven outputs can erode trust, undermine controls, and harm millions of customers. Risk managers face the daunting challenge of ensuring that AI solutions are appropriately governed, risk-aware, and accountable.

The Reserve Bank of India has been monitoring the adoption of AI and ML solutions across regulated entities for several years, particularly since lenders first began deploying ML models for credit scoring. Institutions have been encouraged to develop policies governing the adoption of AI, including maintaining inventories of use cases, tiering by risk materiality, and assigning clear responsibilities for oversight. In August 2024, draft governance guidance was issued specifically addressing models used in credit decisions.

Recognising that credit scoring is only one among many expanding applications of AI, the RBI constituted a committee in December 2024 to examine the broader opportunities and risks of AI adoption. This committee — comprising domain experts, industry practitioners, and RBI officials — was tasked with recommending a framework to ensure responsible and ethical use of AI in India's financial sector. Its report on the *Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI)* was released in August 2025 and sets out a series of recommendations for policymakers and industry participants alike.

At the heart of the FREE-AI Committee Report are the **Seven Sutras** - a set of guiding principles that provide both a regulatory signal and a philosophical anchor. These Sutras articulate key principles which need to underpin AI adoption to promote innovation and progress in the sector while minimising AIs potential to undermine its stability, fairness, and inclusiveness.

The choice of the term *"Sutra"* (meaning a rule or aphorism or collection of aphorisms as a manual) is particularly apt: much as traditional Sanskrit sutras distill profound truths into simple, memorable phrases, these principles are very much designed to anchor AI governance in a set of concise principles to guide policy makers, institutions and the industry as a whole. In the spirit of leveraging classical wisdom, we might suggest another principle for risk managers — *Auchitya*, referring to propriety or appropriateness; the concept is not about rigid rule-following but more about fitness. This notion of fitness can be applied to many aspects of AI deployment – is the level of complexity and intransparency appropriate for the solution at hand? are the oversight mechanisms appropriate to the risk of the solution?

This RiskCounts paper is structured around two objectives:

1. **Section 1**: We present the Seven Sutras from the Committee's report, along with our own brief summary of implications.

2. **Section 2**: We deep dive on the key elements which, from our own experience, we believe will be most relevant for risk managers to practically deliver on the Sutras in their own organisations. We do this through **4 specific "*essays*"** including on embedding governance across the Three Lines of Defense, evolving Model Risk Management and documentation, hardening Third-Party resilience, and investing in Talent and Culture.

In sum, this paper is designed to be both a reference guide and a strategic roadmap - bridging the Committee's recommendations with actionable insights for the Indian financial services industry and positioning *Risk Managers as the custodians of Responsible AI*.

Section I — The Seven Sutras

The Seven Sutras together form a clear North Star for those seeking to responsibly deploy AI, translating these principles into tangible and actionable practices will require purposeful navigation of the many balancing acts required.
Some of the questions which need to be addressed at the institutional level, largely facilitated by Risk teams include some of the following:

While the committee places substantial onus on policy makers to build data and technology **enablers of trust**, as individual institutions, what are the vectors of erosion of trust under our control, and how can we ensure these risks to trust are actively identified and contained?

Sutra 1: Trust is the Foundation

Trust is non-negotiable and should remain uncompromised. In a sector that safeguards people's money, there can be no compromise on trust. AI systems should enhance and not erode public trust in the financial system. When consciously embedded into the essence of AI systems and not treated as a by-product of compliance, trust can be a powerful catalyst for innovation. It is essential to build trust in AI systems and build trust through AI systems.

Sutra 2: People First

AI should augment human decision-making but defer to human judgment and citizen interest. While AI can help to improve efficiency and outcomes, final authority should rest with humans, who should be able to override AI, especially for societal benefit and human safety. Citizens should be made aware of AI-generated content and be informed when interacting with AI systems. Keeping human safety and interest at the core makes AI trusted.

In **putting people first**, what does this mean specifically for our organisation and the clients we serve? How do we ensure that AI delivers empowerment to staff and enhanced experience for customers? Where people serve a safeguard against AI failures, how do we provide the right expertise to our people to be able to effectively challenge AI outcomes?

How do we prioritise **innovation over restraint**, what is our institutional risk appetite in this space? How can we set AI governance whose guardrails are broad enough to enable exploration yet tight enough to detect and prevent unacceptable adverse outcomes? The committee has here only implicitly brought in the notion of notion of **proportionality** – how can risk managers assure that the level of oversight over each case is commensurate with the risks involved?

Sutra 3: Innovation over Restraint

Foster responsible innovation with purpose. AI should serve as a catalyst for augmentation and impactful innovation. Responsible AI innovation, that is aligned with societal values and aims to maximize overall benefit while reducing potential harm, should be actively encouraged. All other things being equal, responsible innovation should be prioritized over cautionary restraint.

Fairness and Equity is perhaps the most daunting challenge for risk managers. AI models have both the potential to 'bake in' historical biases as well as to introduce new ones. But how do we identify, measure and control for fairness? Do we look at disparate impact or disparate outcomes? Over what dimensions should we be testing outcomes?

Sutra 4: Fairness and Equity

AI outcomes should be fair and non-discriminatory. AI systems should be designed and tested to ensure that outcomes are unbiased and do not discriminate against individuals or groups. While AI should uphold fairness, it should not accentuate exclusion and inequity. AI should be leveraged to address financial inclusion and access to financial services for all.

Sutra 5: Accountability

Accountability rests with the entities deploying AI. Entities that deploy AI should be responsible and remain fully accountable for the decisions and outcomes that arise from the use of these systems, regardless of their level of automation or autonomous functioning. Accountability should be clearly assigned. Accountability cannot be delegated to the model and underlying algorithm.

Accountability is clearly set at the institutional level, but what about within the institution? A natural corollary to the 'entity' accountability is the notion of 'user' (ie: business owner of a process) accountability within the entity. But with expertise available to identify and manage risk being split across business, IT, Infosec and Risk teams, how do we ensure that the user has the tools to take on accountability?

The concept of **understandability** (or its close cousin, interpretability) is a discipline still in adolescence – maturing fast but still catching up with the advancements in AI algorithms. We cannot expect the level of transparency we could have in traditional deterministic systems, so the challenge will be to determine how well the outputs of an AI system *can* be understood using the available tools of interpretability, and whether that transparency is adequate is sufficient to confidently deploy the application at hand.

Sutra 6: Understandable by Design

Ensure explainability for trust. Understandability is fundamental to building trust and should be a core design feature, not an afterthought. AI systems must have disclosures, and the outcomes should be understood by the entities deploying them.

For **safety and resilience**, the challenges will be around how to ensure that AI governance frameworks deploy effective risk identification mechanisms that include safety (including information security), and about how to dovetail with operational resilience mechanisms within the organisation. **Energy efficiency**, like understandability is an arena which is evolving rapidly as innovation increasingly shifts from performance accuracy to efficiency. The key challenge here will be how to offset the flexibility and innovation opportunities presented by large, general purpose models against the efficiency of smaller, more special-purpose solutions.

Sutra 7: Safety, Resilience, and Sustainability

AI systems should be secure, resilient and energy efficient. AI systems should operate safely and be resilient to physical, infrastructural, and cyber risks. These systems should have capabilities to detect anomalies and provide early warnings to limit harmful outcomes. AI systems should prioritize energy efficiency and frugality to enable sustainable adoption.

In short, the realisation of effective AI governance is a multifaceted challenge. Leaders in AI adoption already know this, and have been wrestling with these challenges in their own way – it is our hope that with these sutras, the approaches across the industry will coalesce in a way to achieve the systemic and societal goals which the committee envisions.

Section II — Perspectives for Risk Managers

The Seven Sutras recommended by the *Free-AI Committee* provide a foundational framework for the Indian financial system, but their true power lies not in prescriptive language alone - it lies in the philosophical and strategic challenge they represent. **Responsible AI is not a compliance checkbox**; it is a paradigm shift that touches every aspect of banking: risk management, governance, societal trust, and strategic competitiveness. For institutions - and especially risk managers - wishing to harness AI responsibly, the Sutras are a starting point: a call to reimagine both the operational and ethical architecture of finance.

1) Governance and the Three Lines of Defense

Operationalizing the Sutras requires clarity of ownership. The **Three Lines of Defense (LoD)** remains the most practical way to embed that clarity - but each line must be retooled for AI.

First Line (Business & Technology). This is where AI is conceived, built, and deployed. Responsibilities include defining the problem precisely (what outcome or output, for whom, with what constraints), curating data with documented lineage, choosing modeling approaches that are explainable commensurate with risk, and designing human-in-the-loop controls. In our experience, the definition of

responsibility within the first line is also essential – the ultimate owner of risk should be the 'user' - the business function which proposed the use case and owns the process/outcome it serves. Only they know how failure of the AI can impact customers or the institutions, and only they can assess the tradeoffs involved. But these units don't have detailed understanding of the *mechanisms* of failure or the strength or weakness of information security embedded in the solution. For this, they need the support of tech, analytics and infosec teams. To this end, many institutions establish first line risk functions, often resident within the technology function to provide guidance and oversight on controls.

Second Line (Risk & Compliance). The second line must develop a taxonomy of AI risks that includes data quality, representativeness, fairness, drift, stability, privacy, security, explainability, consumer harm, and third-party dependencies. It must set policies and thresholds (e.g., acceptable fairness deltas, maximum drift before review, documentation minimums) and independently challenge first-line assumptions. For LLMs, this includes monitoring for hallucinations, toxic outputs, or compliance breaches in chatbot responses. The second line consolidates results into board-facing dashboards so directors can see outcomes, not just intentions. A key challenge will be to align with the board on a risk appetite for AI-related risk, and alignment on prioritisation of their oversight efforts – risk based tiering which ensures strong oversight on the use cases presenting the greatest vulnerabilities while applying lighter touch were risks where innovation trumps caution.

Third Line (Internal Audit). Internal audit must expand from process compliance to substantive assurance. Auditors need baseline fluency in model life cycles, common explainability techniques, generative AI risks, data lineage, and adversarial testing. They should evaluate whether controls are operating as designed, whether documentation is accurate and complete, and whether the governance culture promotes challenge rather than papering over exceptions. Audit should sample decisions and chatbot outputs back to raw data and prompts to verify traceability — and confirm that escalation paths actually work under load.

Embedding the Sutras across the Lines of Defense transforms principles into day-to-day practice: Trust translates to control efficacy, Accountability becomes named ownership, Understandability becomes documented and testable. Without LoD discipline, even well-written policies remain aspirational; with it, AI governance becomes routine.

2) Model Risk Management and Documentation

Model Risk Management (MRM) discipline is the natural home for AI oversight, but it must evolve. Traditional validation approaches assume relatively stable models. By contrast, AI models are dynamic: they learn on new data, degrade under drift, and may include complex representations that make their decision pathways or outputs non-obvious.

Model Risk Management should begin with a tiered materiality framework. **Materiality is not only financial exposure; it is also societal exposure** — the number of customers impacted, the sensitivity of the output (e.g., product recommendation, fraud block, chatbot guidance), and the difficulty of remediation. Tier-1 models (high materiality) warrant pre-deployment independent validation, conservative rollout, champion–challenger testing, and near-real-time monitoring of performance, drift, and fairness. Tier-2 models use proportionate controls; Tier-3 models still require documentation and basic monitoring but may have lighter oversight.

Validation must expand beyond accuracy to cover the range of adverse impacts relevant for each use case. There is no single minimum test inventory relevant for each model, but some of the tests could include:

- Fairness testing using multiple metrics (e.g., equal opportunity difference, calibration within groups, error rate balance).
- Stability under perturbation (small input changes should not flip outputs).
- Adversarial robustness (guardrails against known exploit patterns).
- Sufficiency of Understandability. As discussed, this is an evolving practice, but a bare minimum should include
    - For predictive models this means use of features supported by business intuition, use of importance tools such as Shapley charts as well as justification for any limits to interpretability
    - For generative models, a structured exploration of the input space to build a good understanding of the relationship between inputs and results

Documentation is the scaffolding of trust. Every material model should carry a model card: problem definition, intended use and limits, key risks identified, detailed methodology (including third party models used), infosec controls and monitoring plan. For predictive models, data sheets should accompany key datasets: provenance, collection method, consent posture, representativeness, known biases, retention limits, and security classification. For chatbots and generative systems, documentation should also specify red-teaming results, content filtering approaches, escalation paths, and categories of prohibited use.

Finally, MRM will need to address lifecycle governance: pre-deployment approvals, staged rollouts with holdouts, periodic revalidation (triggered by data shifts or policy changes), and formal retirement with data and artifact archiving. *When MRM is treated as a living practice rather than a one-time gate, AI becomes governable at scale.*

3) Third-Party and Vendor Risk

Modern AI is built on ecosystems: cloud infrastructure, external datasets, pretrained models, model-as-a-service APIs, and open-source libraries. These bring speed and capability - but also concentration risk, opacity risk, and critically, exposure of sensitive data across organizational boundaries. With accountability resting with the deploying institution, the challenges for risk managers is establishing what visibility into the vendor control environment is sufficient for each use case, and what expertise (technical, analytical and infosec) needs to be brought to the table to give adequate comfort to accept the use case as implemented.

Vendor risk programs must move beyond generic questionnaires to AI-specific due diligence. Contracts should require:

- Transparency about model lineage, training data categories, update cadence, and known limitations.
- Notification and approval before material model changes go live.
- Right to audit (including secure model review or third-party attestation).
- Service-level objectives for fairness monitoring, uptime of safety controls, and incident response.
- Explicit commitments on data residency, privacy, and security, including encryption, access management, and localization.
- Exit and portability provisions, including a plan for continuity if the vendor fails or withdraws service.

For open-source components, institutions should maintain a software bill of materials (SBOM) and track vulnerabilities and licenses that may affect usage rights or security.

Operational resilience requires asking a set of questions about dependence on vended AI processes above those part of more conventional resilience considerations. What happens if a major foundation model pushes an update whose behaviour deviates from prior expectations? Do we have a fallback model, a rule-based backstop, or a degradation mode that maintains safe service with reduced functionality? Are rate-limiters in place to prevent vendor errors from impacting delivery of services?

Risk managers must emphasize that accountability is not transferrable. Institutions can outsource technology, but not responsibility. Consumers and regulators will hold the deploying bank accountable—not the vendor. Governance, monitoring, and remediation must therefore remain in-house competencies, even when models are bought rather than built.

4) Talent, Culture, and Human Judgment

Technology cannot substitute for ethics or judgment. Responsible AI depends on people who are skilled, empowered, and rewarded for doing the right thing ... even when it is less expedient; slower or less convenient.

Institutions should design a capability architecture for AI governance. Beyond data scientists and engineers, you need product owners who can frame ethical boundaries, risk managers who understand drift and bias, internal auditors with model literacy, and frontline staff trained to interpret and challenge AI outputs. Training will need to be role-based: call-center and branch teams need to recognize when to escalate; product and credit teams need to understand fairness metrics and consumer protection implications; executives and directors need to read AI dashboards with skepticism and ask for evidence, not reassurance.

Culture is the multiplier. A speak-up environment is not a slogan; it is a set of practices: anonymous channels, non-retaliation guarantees, and performance systems that value escalation and learning. Overrides should be treated as signals, not as mistakes to hide. Patterns of override — by product, geography, customer segment — should feed back into model improvement and policy reconsideration.

Human-in-the-loop is credible only if humans have time, tools, and authority. That means understanding of the models, user interfaces that surface clear explanations, not probability scores alone; it means escalation SLAs so customers are not trapped in automated limbo; it means governance rituals (e.g., monthly fairness councils, model risk committees) where issues are reviewed across functions.

In the end, the Sutra "People First" lives or dies in culture: when employees know they are expected - and protected - to use judgment over automation.

Conclusion

We see these as the four key pillars which risk managers need at their disposal – Three Lines, evolving Model Risk Management, hardening Third-Party Resilience, investing in Talent and Culture - together these shift *Responsible AI* from a compliance exercise to a strategic capability: one that protects customers, strengthens stability, and earns the trust that enables innovation to scale.

The Seven Sutras, then, are not simply guiding aphorisms. They are an invitation to rethink the philosophy of finance in an AI-driven world - reminding us that trust, fairness, accountability, explainability, and resilience are the very conditions for sustainable financial innovation. Done

right, Responsible AI becomes a strategic differentiator and risk functions become enablers of confidence: earning public trust, attracting customers, satisfying regulators, and building resilience into the financial system.

The challenge is immense, but so too is the opportunity. India's financial sector is uniquely positioned to demonstrate to the world that Responsible AI is not a theoretical aspiration but a practical, strategic, and ethical reality. If the Sutras are embedded not only in regulation but in the DNA of institutions, India can lead by example - showing that the future of finance is both digital and humane.

About RiskCounts

RiskCounts is a niche Risk Management and Compliance Advisory firm, dedicated to helping businesses navigate the complexities of regulatory requirements and risk mitigation. Our services are designed to empower organizations to identify, assess, and manage risks while ensuring full compliance with industry standards and regulations. We offer tailored solutions in areas such as regulatory compliance, risk assessments, internal controls, and corporate governance. With our deep expertise and hands-on approach, we work closely with clients to develop strategies that not only minimize risks but also enhance operational resilience and foster sustainable growth. At RiskCounts, we are committed to being your trusted partner in building a strong compliance framework and a risk-aware culture that drives long-term success.