Responsible Artificial Intelligence (AI) – Balancing Innovation with Financial Stability[1]

Opening and Context Setting

Good afternoon, distinguished policymakers, members of academia, industry leaders and innovators. It is both a pleasure and a responsibility to address this gathering on a subject that is poised to shape the future of finance, society, and governance alike—*Responsible Artificial Intelligence*.

AI has rapidly evolved from an academic discussion less than a decade back to become an integral part of our daily lives. We encounter it when we unlock our phones, interact with chatbots, and increasingly, when accessing financial services. In just a few years, AI has evolved from an enabling technology to a foundational driver of how individuals and businesses make decisions.

Globally, AI is already reshaping financial systems. From digital credit underwriting to conversational banking assistants, AI is demonstrating its ability in ways unimaginable a decade ago. India, too, has been a notable participant in this journey.

Yet, as with all powerful innovations, AI carries a dual narrative. It promises extraordinary efficiency, inclusion, and innovation, but if left unattended, could pose unprecedented threats. As Stephen Hawking said in 2016 at the launch of the Centre for the Future of Intelligence (CFI), “the rise of powerful AI will either be the best or the worst thing ever to happen to humanity. We do not yet know which.” It is acknowledged widely that AI could be the permanent answer to poverty and disease. But equally there have been concerns from AI experts – ranging from concern around bad actors using AI for bad things to the more fundamental concern that human existence is irrelevant once machines achieve superintelligence. I do not intend to dwell on these widely divergent possibilities but only to highlight the limited point that while the benefits of AI are transformative, they need to be used responsibly. In finance, the margin for error is even narrower as financial institutions are built on trust and economies prosper on stability. Therefore, the integration of AI in financial systems must be approached as a matter of profound responsibility with due recognition and mitigation of risks.

[1] Keynote address delivered by Deputy Governor T Rabi Sankar at the Global Fintech Festival, Mumbai on October 7th, 2025

AI for the Financial Sector

The Benefits

The promise of Artificial Intelligence in finance is by now well recognised. At its core, AI can expand financial access, strengthen safeguards, and reimagine efficiency. It can lead to better credit assessment through use of alternative data (like transaction patterns, utility payments, etc.) of unbanked customers. Ability to use massive data sources could help in real-time detection of frauds through identification of unusual transaction patterns, or improve market risk modeling.

Operational efficiency and cost reduction can get a paradigm shift using AI, e.g., in back-office processes, KYC, loan processing etc. Chatbots and virtual assistants would achieve 24x7 customer support. Data extraction from financial documents (e.g., invoices, contracts) through Natural Language Processing (NLP) could make document processing seamless.

In the investment and trading space, the ability of AI models to detect short-term price inefficiencies are already being harnessed. Other benefits include allocation optimisation, use of big data to forecast market movements etc. AI driven RegTech applications help regulated entities with better compliance outcomes. Automated monitoring helps detection of suspicious transactions and generates faster compliance.

AI has the potential to significantly expedite financial inclusion through alternative credit scoring models while language interfaces will remove language barriers and reach digitally limited customers. Investment advice becomes affordable to small investors through robo-advisors.

The Risks

But these benefits come with significant risks. AI systems are trained on vast amounts of data. It is natural that the learning from the data would also extend to learning the bias inherent in data. AI systems trained on biased historical data are likely to perpetuate or amplify historical discrimination in, for example, credit profiling, or hiring. Even small biases in training data can lead to systematic exclusion of population groups from accessing financial services. Algorithmic opacity would make it difficult to identify possible biases.

The ‘Black Box’ problem of AI models, or, in other words, the lack of explainability, makes these models non-transparent. This makes it hard for regulators and auditors to understand how decisions are made, which, in turn, undermines accountability. Regulatory actions, or denial of service to customers, would typically require reasons to be communicated. Absence of explainability may thus constrain the use of such tools.

There are systemic risks typical to AI systems, such as herding behaviour when AI-driven trading models get widely used, which can amplify volatility. AI misjudgments can trigger market dislocations. Such problems are amplified by the possibility that it becomes difficult to assign responsibility when an AI makes a harmful or erroneous decision. Legal frameworks would always find it difficult to catch up with fast moving technology. Over-reliance on automation could result in losing oversight or delayed intervention when things go wrong. There is also the ethical issue of using behavioral data for manipulative cross-selling or risk profiling.

Then there is the ongoing debate about AI and job displacement. Whether AI will displace jobs in the long run would depend on whether it is like other transformative changes in history like the Industrial Revolution or the invention of electricity or whether it is a fundamentally different kind of change. Max Tegmark, founder of Future of Life Institute argues that while all past technologies amplified human ability but did not replace human intelligence, AI is the first technology that creates intelligence itself.

Recognizing these risks is not to diminish the promise of AI, but to underline the importance of adopting it responsibly through safeguards, governance, and foresight.

Balancing Innovation with Stability

The key question is, how do we enable innovation while safeguarding systemic stability? This balance is a necessity for ensuring that AI strengthens rather than undermines the financial system. If regulatory frameworks are too rigid, they can dissuade experimentation, reducing AI to a tool deployed only by the largest players. On the other side, unbridled adoption, particularly in high-impact areas, could create vulnerabilities that are invisible until they snowball into crises.

The balance is not only about restraint; it is also about actively encouraging innovation. This requires policies that create safe spaces for experimentation, such as sandboxes, facilitate open digital infrastructures, and provide access to quality data, enabling firms to innovate with confidence. It also requires incentives for responsible innovation, so that firms see governance not as a burden but as a competitive advantage.

Responsible AI – Guiding Principles

As we reflect on the transformative potential of AI, it becomes imperative to anchor its adoption within a framework of principles. The RBI took a proactive step through setting up the FREE-AI Committee, which has articulated a set of guiding *sutras* for responsible and ethical adoption of AI in the financial sector. These principles are intended to serve as touchstones for all stakeholders.

At the core is the principle of trust, the bedrock of finance. Every deployment of AI must reinforce, not diminish, the trust of consumers, institutions, and society. Equally important is a people-first orientation, ensuring that technology serves human needs. The report emphasizes innovation over restraint, coupled with fairness and accountability in outcomes. AI can inform decisions, but it cannot own them. The accountability must always rest with human actors and institutions deploying AI.

The principle of ‘understandable by design’ underscores the need for transparency, ensuring that AI decisions are explainable to both regulators and consumers. And above all, safety and resilience must be built into every layer of adoption.

Alongside regulatory oversight, it is equally critical to encourage industry-led codes of conduct, self-regulation, and the institutionalisation of ethical standards. This collaborative approach ensures that responsibility is not a mandate of regulators but a shared culture across the fintech ecosystem.

RBI’s Approach and Role

The RBI has always fostered “innovation within safeguards.” Through calibrated guidance, supervisory oversight, and structured engagement with industry, the RBI aims to foster an ecosystem where financial innovation flourishes without compromising systemic stability. As AI reshapes the financial landscape, this approach remains unchanged - progress and prudence must go hand in hand.

RBI has also taken initiatives for the industry through RBIH, such as MuleHunter.ai™ for combating the menace of mule accounts. Unlike the traditional rule-based systems currently used by banks, MuleHunter.ai™ offers greater accuracy and precision with significantly low false positive rates. Currently, the model has been deployed in about 20 commercial banks. In addition, work is also underway to explore a Digital Payments

Intelligence Platform (DPIP), that can analyse and assign a risk score to transactions on a real time basis.

Ringfencing and Guardrails

While AI holds immense promise, the financial system demands the highest degree of prudence. Critical infrastructures and institutions must be ringfenced from unchecked risks that could arise from untested or poorly governed AI deployments. The objective is not to obstruct innovation but to ensure that its application never compromises the stability or integrity of the system.

To this end, practices such as stress-testing of AI models under diverse scenarios, red-teaming to identify vulnerabilities, and the adoption of explainability tools and standards are indispensable. These mechanisms would help regulators and institutions alike to supervise AI outcomes, detect weaknesses before they escalate, and ensure that AI-driven decisioning can be understood and, if necessary, challenged.

Equally important is that AI systems are subjected to rigorous oversight and layered with inherent checks. Financial AI applications must be designed such that they cannot inadvertently destabilize markets, payment systems, or consumer confidence.

This approach demands “safety by design” rather “safety as an afterthought.” Safeguards must be embedded throughout the lifecycle, from conception and data training to model validation and real-world application. Retrofitting safety once risks have materialized is inadequate and potentially destabilizing.

Research, Innovation, and Collaboration

Embedding AI in finance is not a one-time exercise. It demands continuous research, experimentation, and learning, as models, techniques and risks evolve rapidly. It also calls for partnerships among industry, academia, regulators, and start-ups. Co-developing solutions, sharing knowledge, and stress-testing innovations must be fostered.

Entities should have in place systems for responsible data governance, ethical sourcing of data, and privacy-by-design in every model. They should develop common standards, toolkits, and disclosure mechanisms so that model design, training data, and decision logic can be explained to regulators and customers. Safeguards such as digital watermarking of synthetic content should be explored to deter misuse. Internal policies and processes must be revised to embed AI risk assessment into the product lifecycle. Continuous monitoring, stress scenarios, and independent audits should be institutionalised.

The Way Forward

As we look ahead, the path for responsible AI in India's financial sector is exciting yet deliberate, demanding a phased approach that balances innovation, inclusion and stability.

Alongside technological progress, the human element remains central. AI literacy for consumers to understand both the potential and risks of AI will be critical. Just as financial literacy has been a national priority, the coming decade will require a parallel focus on AI literacy, for individuals to engage confidently and safely with these new tools.

In the short term, the focus needs to be on awareness and capacity building. Financial institutions, technology providers, and regulators must train personnel, strengthen internal governance structures, and introduce initial risk frameworks to ensure that AI deployment is encouraged with focus on safety. Awareness campaigns and workshops can also help smaller institutions and FinTechs integrate AI responsibly.

In the medium term, the FREE-AI principles should guide the industry practice. AI can begin to play a substantial role in SupTech, credit decisioning, and financial inclusion. In parallel, the industry should develop its own governance standards, self-regulatory codes, and ethical guidelines to complement regulatory oversight.

In the long term, India can aspire to become a trusted global hub for responsible AI in finance. By demonstrating how innovation can coexist with strong safeguards, India can set an example for emerging economies and the Global South, attracting talent, investment, and collaboration.

Concluding Remarks

As we conclude, it is essential to reiterate that AI must remain a force for good - empowering individuals, strengthening institutions, and enhancing the resilience of our financial system. Its promise will be realised only when adopted responsibly, with constant attention to societal impact.

Responsible AI should not be framed merely as a regulatory requirement, but as a matter of business ethics. Every model deployed, every decision automated, and every service enabled through AI must reinforce the confidence of consumers, provide fair access, and respect the dignity and privacy of all participants.

Let me leave you with five guideposts. Not as tasks, but as a collective mission: the 5Ts

1. **Trust** – Commit to building AI systems that uphold and enhance the trust in the system. Embed responsibility and ethics in every algorithm.
2. **Transparency** – Reinforce clarity and explainability in AI, ensuring that decisions can be understood, audited, and questioned when necessary.
3. **Training** – Invest in training to nurture a world-class AI talent within our financial ecosystem. Ensure India leads in creating, not just consuming.
4. **Technology for Good** – Let innovation be guided by purpose. The test of every AI application must be whether it advances inclusion, resilience, and efficiency.
5. **Togetherness** – Above all, we must work together. Regulators, industry, academia, and global partners, to collaborate and co-develop.

Through shared commitment, ethical deployment, and continuous vigilance, we can ensure that AI fulfils its promise as a transformative enabler.

******