pwc

DVARA Research

FACE
The RBI-recognised Self-regulatory Organisation in the FinTech Sector (SRO-FT)

# Principles of responsible and trustworthy AI in digital lending

March 2026

# Foreword

**Vivek Belgavi**
Partner, Financial Services Advisory Leader
PwC India

Trust is the cornerstone of financial institutions. With AI taking on an increasingly prevalent role in the digital lending space, trust must remain at the centre of how we develop and deploy these systems. Since AI has the potential to speed up the processes and reduce turnaround times while bringing millions of underserved borrowers into the formal financial system, for developing nations like India, AI adoption presents an opportunity to reduce the credit gap and drive meaningful change in the financial lending ecosystem.

However, this opportunity also brings new risks and challenges. Automated credit decisions reflect the data they are trained on, the objectives they are optimised for, and the choices made by those who build and oversee the AI systems. As a result, biases in AI models, decision-making, and governance protocols can erode consumer trust, attract regulatory scrutiny, and undermine the credibility and viability of the institutions which deploy them. Therefore, responsible and trustworthy AI is a strategic imperative for leaders in financial services.

The need for employing responsible AI practices doesn't simply lie in risk management and regulatory compliance. Lenders who'll invest in transparent and ethical AI systems will be better positioned to build durable customer relationships and earn the trust of regulators and the broader ecosystem which could lead to commercial gains in the long run.

We are proud to collaborate with Dvara Research Foundation and the FinTech Association for Consumer Empowerment in developing this paper which offers a grounded and actionable perspective on embedding RAI principles across the digital lending value chain, supported by a practical framework and a self-assessment tool to help institutions translate RAI commitments into concrete actionable steps. We hope this paper serves as a valuable resource to our readers.

# Foreword

**Rajnil Malik**
Partner, AI GTM Leader
PwC India

AI has moved rapidly from being a concept to a core driver of business operations and innovation. In digital lending, AI-powered solutions are reshaping how credit is assessed and delivered, enabling faster decisions, enhancing customer experience, and extending financial access to the underserved. However, as AI becomes deeply embedded in lending operations, it becomes essential to minimise the consequences of weak design, inadequate governance, or misuse and look after safeguarding the interest of both consumers as well as financial institutions.

This makes responsible AI (RAI) essential. In digital lending, where automated decisions can impact financial outcomes, principles such as fairness, accountability, and trust must be built into AI systems from the very beginning. This urgency is reflected in the Reserve Bank of India's Free AI Report, which reinforces many of the themes explored in this paper, and highlights transparency, robust risk management, and consumer protection as key pillars of a trustworthy AI ecosystem in financial services.

Navigating these challenges requires more than technical expertise. The ethical use of AI in digital lending needs a multidisciplinary approach, one which balances business objectives, technological advancements with regulatory expectations and most importantly, the rights and lived experiences of borrowers.

PwC is glad to collaborate with Dvara Research Foundation and the FinTech Association for Consumer Empowerment for developing this paper which presents a comprehensive view of how AI is currently being deployed across the digital lending value chain, supported by practical and actionable steps that lenders can incorporate into their AI systems to operationalise RAI principles. To further support the implementation of RAI principles, the paper also includes a checklist and self-assessment tool that organisations can use to evaluate and strengthen their adherence to RAI practices.

As the digital lending landscape continues to evolve, we believe that organisations that lead with RAI will be better positioned to build trust with customers, regulators, and the broader ecosystem. We are pleased to share these insights and to support our clients and partners in advancing AI systems that are transparent, responsible, and sustainable.

# Foreword

**Indradeep Ghosh**
Executive Director
Dvara Research Foundation

AI is increasingly being integrated in the design and delivery of financial services. A survey conducted by RBI's Department of Supervision found that about a fifth (20.8%) of the supervised lenders report using AI in their operations and an even larger proportion expresses an interest in incorporating AI.[1] Stakeholders in the financial services sector, academia and policymakers collectively acknowledge the importance of implementing responsible and trustworthy AI practices as they understand that reaping benefits of AI and reducing risk of attendant harms requires careful and deliberate actions.

The Indian policy discourse, first through the RBI's FREE-AI Committee and then through the India AI Governance Guidelines, has presented a rich framework of 'RBI's seven sutras' to ensure that AI can be used in a responsible and trustworthy manner. The sutras are universal guiding principles for making AI transparent, trustworthy, reliable, and context agnostic.

The natural next step for each sector is to contextualise these sutras and operationalise them through actionable practices and this whitepaper is a step in that direction. The paper also contains a web-based checklist of best practices that allows digital lenders to align their use of AI in conformity with the mandate of responsible and trustworthy AI (RTAI). The checklist is a diagnostic and corrective tool. It scores the strength of the existing AI safeguards of digital lenders, allowing them to gauge the gaps in implementing the recommended best practices. It also highlights aspects that require urgent attention, allowing lenders to recalibrate their practices and achieve greater alignment with RTAI principles. In doing so, the checklist translates universal principles into well-defined processes and bridges the gap between principles and practices. It is designed with industry participants and comprises practices that are considered necessary for operationalising RTAI globally.

We hope that you find the paper insightful.

1 https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF

# Foreword

**Sugandh Saxena**
Chief Executive Officer
FinTech Association for Consumer Empowerment (FACE)

From customer screening to credit-risk assessment, and from identifying fraud to pricing, servicing and collection, automated decisions bolster many elements in the digital lending lifecycle. Since AI is regularly retrained to work with third-party vendors and human oversight, no single source can be held accountable for how final credit decisions are produced and/or reviewed. Often, this leads to a scenario where institutions have limited visibility on their decision origin, and material gaps in accountability arise.

The RBI regularly attempts to develop frameworks to mitigate these issues, including the recent FREE-AI Committee Report which presents the guiding principles of the seven sutras and recommendations for shared AI use and oversight. An important aspect of the FREE-AI report is the mention of SROs such as FACE, envisioned to act as intermediaries to convert AI principles into operational standards, norms, and checklists that the industry can employ for governance and peer-benchmarking.

The same principles of the report have also been discussed in this white paper, though its innate value lies in drawing attention to upstream decisions. Since the paper examines AI systems across the model's lifecycle, lenders now have a structured way to locate responsibility among complex tech-stacks and distributed teams. The paper supplements this identification with concrete practices to translate responsible and trustworthy AI (RTAI) from start (solution design, and model development) to finish (deployment, and monitoring). The paper also offers recommendations to refine AI not only at the development stage (e.g. stakeholder input in model design) but also for implementing RTAI's principles (e.g. establishing ethical review boards and undertaking fairness assessments).

FACE regularly engages with its membership base of 300+ FinTechs, and understands that FinTechs cannot have uniform capability or intent. Thus, the paper's distance mapping approach is beneficial for FinTechs to assess their current practices across the various stages of AI adoption and maturity.

As AI is increasingly being used in digital lending to scale its operations and enhance customer satisfaction, this whitepaper and the self-assessment toolkit could play a significant role in ensuring that AI solutions can be implemented in alignment with RTAI's principles. We hope that you find this paper to be useful and insightful.

# Contents

# Introduction

The adoption of artificial intelligence is steadily increasing across various industries. The financial sector, in particular, is at the forefront of this movement, and is harnessing AI to realise considerable commercial value while simultaneously enhancing operational effectiveness, enriching customer experience, strengthening risk management frameworks, and promoting innovation. Some of the benefits of AI adoption in FinTech are:

1. **Data processing abilities:** Enhanced data processing abilities, including the ability to process qualitative and audio-based data could lead to a deeper understanding of customers' needs and improve product fitment. These insights could also significantly improve the customers' journey by sensing their needs and offering customised, relevant and timely support in a customer-friendly format along with the customer journey. AI systems can also help financial service providers (FSPs) build stronger defenses against fraudulent activity.
2. **Flexibility and scalability:** AI systems exhibit a high degree of scalability and flexibility which enables FinTech organisations to offer hyper-personalisation at scale.
3. **Process-rationalisation:** AI systems help realise efficiency gains from process-rationalisation where providers can eliminate duplication of tasks by deploying AI.

When AI works as intended, it can deepen financial inclusion and enhance the relevance of financial services at population-scale. For instance, GenAI can significantly enhance the ease of opening accounts for uninitiated customers. It can also nudge them to improve account usage, promote budgeting, and deepen financial literacy through relevant, customised and timely content. In the case of credit, fuelled by big data, algorithms could do a better job in predicting creditworthiness of thin-filed customers and credit invisibles. These gains, however, are tempered by attendant risks. AI-related risks could arise from the advanced processing of rich and personal data. This includes risks related to data privacy, bias and discrimination, AI hallucination and misinformation, and inconsistent accuracy of the AI system. These risks can take away from the gains presented by deeper processing capabilities. Further, these risks are aggravated by the relative scalability of algorithms.

Just like the benefits, the risks could also increase with the model, affecting very large number of customers at once. The difficulties in explaining AI processes and implementing complex algorithms could make the algorithm impermeable and difficult to assess for accuracy, thereby reducing the scope for customers to question the algorithm, identify mistakes or seek remedial action. These risks can trigger adverse systemic shifts in the financial system and jeopardise customer safety. For instance, reliance on similar, off-the-shelf machine learning tools could encourage herd behaviour among lenders which could intensify economic volatility. Similarly, a less-than-fit algorithm could enable lending to those borrowers who may not have the wherewithal to repay the loans. It could affect the borrowers' credit scores, cutting them off formal credit markets and severely erode the lenders' portfolio quality.

Assessing and understanding the risks and benefits of using AI is essential for drafting informed policies on how AI can be integrated in the product development process. This must be supplemented with a detailed analysis of the assessments' underlying causal mechanisms. AI governance must adopt a lifecycle or a value-chain approach which focuses on improving the visibility of the various components of an AI system and the stakeholders responsible for each. It allows organisations to identify the origins of the risks as well as the benefits along the value chain and allocate responsibilities accordingly. This visibility over the various components, their potential implications and the respective stewards also benefits the regulators.

Finally, the value-chain approach to governance is necessary for crafting an AI governance framework that focusses on responsible and trustworthy AI (RTAI). RTAI concerns itself with the AI system as-a-whole and not just the outcomes of AI adoption. It also requires AI systems to be technologically robust and aligned with socially desirable values such as non-discrimination and fairness to minimise biases and any instances of data breach.

Though responsible and trustworthy are often used interchangeably, this paper distinguishes the two as follows:

a. Responsible pertains to the processes associated with the design and deployment of AI

b. The conduct and the outcomes of the AI system thus designed fall under the ambit of trustworthy.

Thus, 'responsible' describes the processes and procedures put in place to ensure that the conduct, decision and outcomes of the AI systems are trustworthy.

The paper explores what RTAI means in the context of digital lending. The first section compiles principles of RTAI along with its essential components. The next section maps relevant tools for each principle. These tool recommendations can help lenders implement RTAI practices in their operations. The distance map—a checklist for technology teams of digital lenders—allows digital lenders to gauge how far their current AI safeguards are from the desired level and how they might close this gap. The teams should be able to review the checklist without external supervision and reflect on the overall intensity of their AI safeguards. The map serves as a diagnostic tool and offers guidance to lenders on how they might further strengthen their AI practices.

The distance map is anchored in the understanding that AI in finance needs to be governed not for AI's sake but for the sake of finance. The recommendations of the checklist prioritise mitigating the risks of the indiscriminate use of AI. Since these recommendations are compatible with the legacy systems and current processes of digital lenders, they can be implemented at minimal cost and provide the benefits of responsible AI practices in digital lending.

# 1 Principles of RTAI

Responsible AI aims to ensure that AI systems function as intended and uphold fair practices while minimising risks. The adoption of responsible AI ensures that AI-driven decisions and outcomes are trustworthy for those directly impacted by them. It is more than just a regulatory checkbox and has implications for overall customer protection as well as the stability of the broader financial system. For AI to be responsible it must exhibit following characteristics:

## 1. Transparency, explainability and contestability

Transparency provides stakeholders with a contextual overview of the workings of the AI system. It includes the ability to trace the origins of the data, clearly identify automated decisions and understand the limitations of the AI system. At the same time, the principles allow the developers and deployers of AI systems to safeguard their intellectual property and trade secrets. For this principle to be effective, it must provide information that is appropriate for the actor seeking it out. While transparency does not guarantee accuracy, it makes it more likely by enabling an enquiry into the logic of the system and the origin of the underlying data.[2]

Explainability means enabling people to understand how an AI system was developed. This entails providing easy-to-understand information, which empowers those directly affected by the outcome of the AI to challenge it. The principle comes with reasonable restrictions on how it is meant to be applied. It calls for AI developers to ensure that the output is understandable by sharing the underlying factors and logic (exogenous explainability), without having to disclose the intricate details of the model itself (decompositional explainability). These restrictions help protect the model from adversarial attacks while preserving valuable proprietary information that maintains its competitive advantage.

Contestability enables timely human review and remedial action if an automated system fails or produces errors, especially in sensitive contexts. For instance, women-owned businesses have historically faced higher loan rejection rates. A credit scoring model trained on this data may associate this demographic with higher risk and deny them loans, thereby reinforcing the bias.

Transparency, explainability and contestability obligations may differ for digital lenders based on their use cases.

2 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Globally, these attributes are being determined by the sensitivity of the use case and the implications of a misjudgement on the part of the AI (EU AI Act). Excessive transparency could create confusion or expose the AI models to exploitation or manipulation. Regulators such as the Monetary Authority of Singapore (MAS) recommend that the sophistication of the explanation should match the expertise of the agent querying it. Similarly, excessive explainability could incentivise developers to reduce the number of variables in the model, reducing its accuracy.[3]

## 2. Fairness and non-discrimination

The principle of fairness and non-discrimination mandates that AI systems should be designed and implemented to prevent biases and discriminatory outcomes by fostering inclusivity, transparency and regular monitoring throughout the AI lifecycle while also complying with legal and ethical standards to protect against any form of discriminatory outcomes. For example, discrimination here would entail dissimilar credit terms to individuals who are similar in their creditworthiness. Some regulators have set out practical guidance for implementing fairness which emphasises that no two groups or individuals should be treated differently without justification and the justifications provided should also be frequently reviewed for accuracy.

3 https://oecd.ai/en/dashboards/ai-principles/P7

## 3. Technological dependability of the AI system (and its ability to respond to realised risks)

This attribute comprises three characteristics:

**(i) Reliability:** Reliability of the output is the ability of an AI system to perform as intended, without failure, for a given time interval, under given conditions. In its essence, reliability ensures that an AI system behaves exactly as its designers intended and anticipated and repeatedly exhibits the same behaviour under similar conditions. Reliability is a goal for overall correctness of the AI system.[4]

**(ii) Robustness and resilience:** Robustness, in the context of AI systems, refers to the ability of an algorithm or a model to maintain its accuracy and stability under different conditions, including variations in input data, environmental changes and attempts at adversarial interference. It ensures that the system can withstand unforeseen challenges and continue to function effectively. Resilience of AI refers to its ability to bounce back after disruptions.

**(iii) Safety and security:** The principle of safety refers to reducing 'unintended' behaviour in the functioning of AI. It aims to prevent unwanted harms to human life, health, property or environment.[5] The principle of security requires robust systems to be installed which can defend the AI systems against security risks and to respond to and recover from these risks. Security risks in AI include concerns related to the confidentiality, integrity and availability of the system and its training and output data. Further, a secure system can maintain the integrity of the information that constitutes it. This includes protecting its architecture from unauthorised modification or damage of any of its parts. A secure system continues to operate reliably and remain available to authorised users, while protecting sensitive and private information, even when it is exposed to hostile or adversarial conditions.

## 4. Privacy and data protection

The principle of privacy and data protection seeks to safeguard individuals' privacy rights by requiring the collection, processing and storage of data to be conducted with transparency and consent, adhering to relevant laws and ethical standards. It also emphasises the importance of minimising data usage, ensuring that data processing does not reveal sensitive information which is not relevant to the context and which the data principal would not have ordinarily shared with the business. Further, it emphasises ensuring data security and providing individuals with control over their personal information, including rights to access, correct and delete their data.

4 https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

5 https://www.nist.gov/artificial-intelligence/ai-fundamental-research-security

### 5. Protecting human agency and establishing human oversight

This principle seeks to preserve human autonomy by ensuring that the AI system remains accountable to a human agent. It also emphasises that individuals should be able to make informed and autonomous decisions regarding AI systems. This can be achieved by implementing human oversight in the functioning of the AI system.

### 6. Governance and accountability

Governance underscores the need for effective frameworks for AI systems' development, deployment and use in accordance with the rules set by external bodies and internal governing teams within the organisation. It includes risk management, regular assessments and stakeholder involvement, and mandates reporting to allocated boards for responsible AI's adoption.

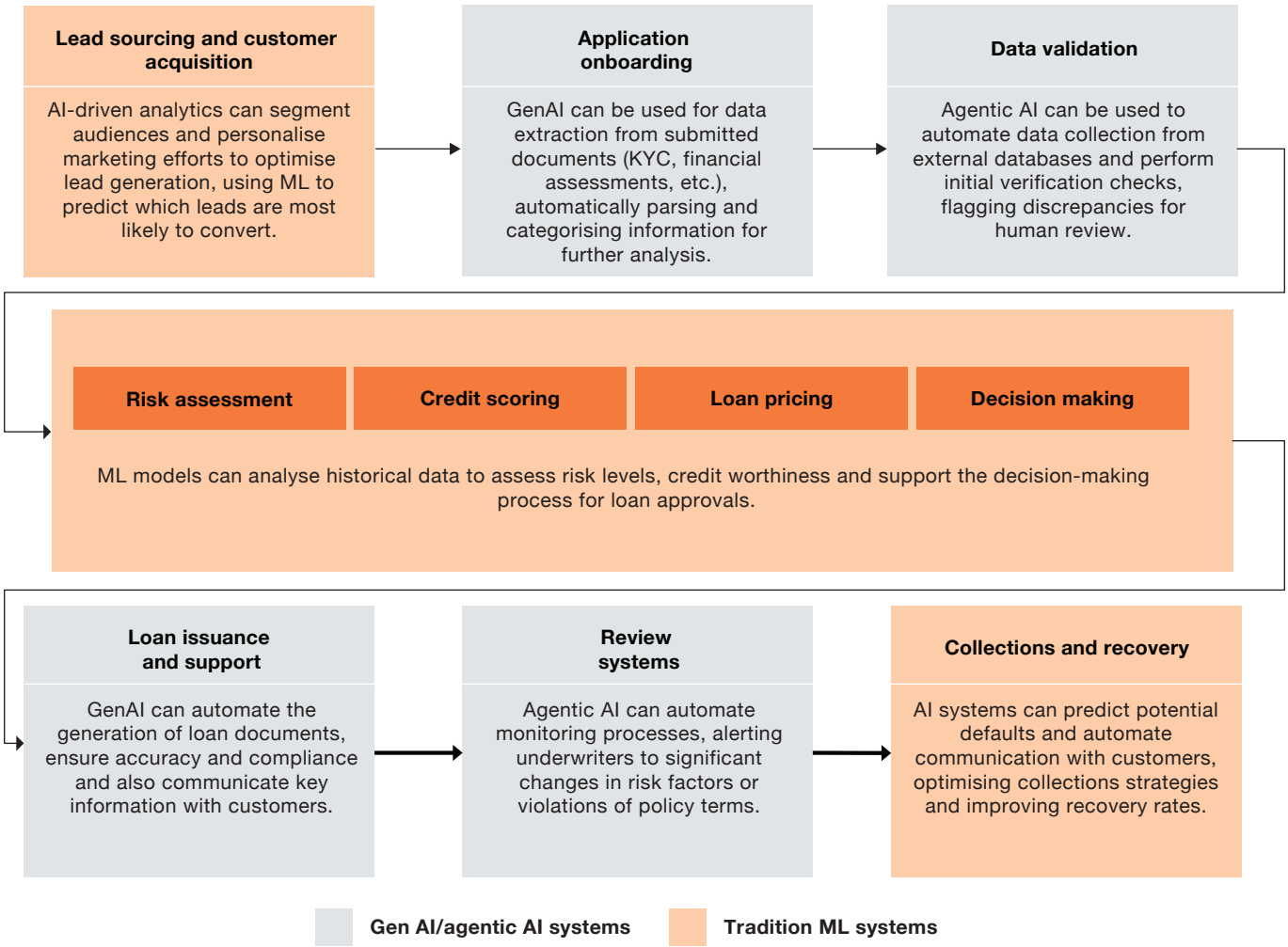
Accountability ensures that those responsible for the development and deployment of AI systems are able to demonstrate their adherence to the principles and practices of RTAI. This principle ensures that responsible actors can be held accountable for their roles in the functioning of the AI system.

# 2 How RTAI can help digital lending

From the use of predictive analysis to understand customer behaviour during the acquisition stage to AI-powered chatbots and virtual assistants enhancing customer support, AI is already reshaping the ways in which digital lenders operate. AI is also being used in core business functions including credit scoring models trained on traditional or alternative data and personalisation of terms and conditions for borrowers based on their profiles. The governance of AI in digital lending is crucial for making digital lending safe for customers and lenders. This section discusses the use of AI by digital lenders. It also highlights the key concerns that arise in each phase of designing the AI system, and the RTAI practices.

## Mapping the use of AI in digital lending

**Figure 1:** Key stages of a digital lending value chain




Note: The use of AI systems mentioned in the diagram above is indicative/interchangeable and has been tagged in this manner to illustrate the potential uses of these different systems across the digital lending process.

## Model development lifecycle (MDLC) for risks and gains for lenders

MDLC is the structured process for building, deploying and maintaining AI models. The process comprises a series of steps such as defining the problem, preparing the data, model designing and training, and performance monitoring. Robust MDLCs incorporate several processes and a continuous feedback loop to ensure that the model is accurate, reliable and scalable. A typical MDLC comprises three essential phases

(i) **Solution design phase:** This phase involves defining the business requirements, identifying problems, selecting the appropriate methodology and designing the overall architecture and workflow for the model. It also includes preparing the data required for the model, including identifying the sources of relevant data, cleaning it and structuring it for use.

(ii) **Model development phase:** In this phase, data is collected, processed and analysed to build, train and validate the model using selected algorithms and techniques.

(iii) **Model deployment and monitoring phase:** This phase focuses on integrating the model into production systems, making it operational, and continuously monitoring its performance for all the RTAI principles.

Juxtaposing this understanding of the MDLC with the risks and gains associated with the use of AI in lending allows users to understand the vulnerabilities and benefits for drafting the right RTAI model.

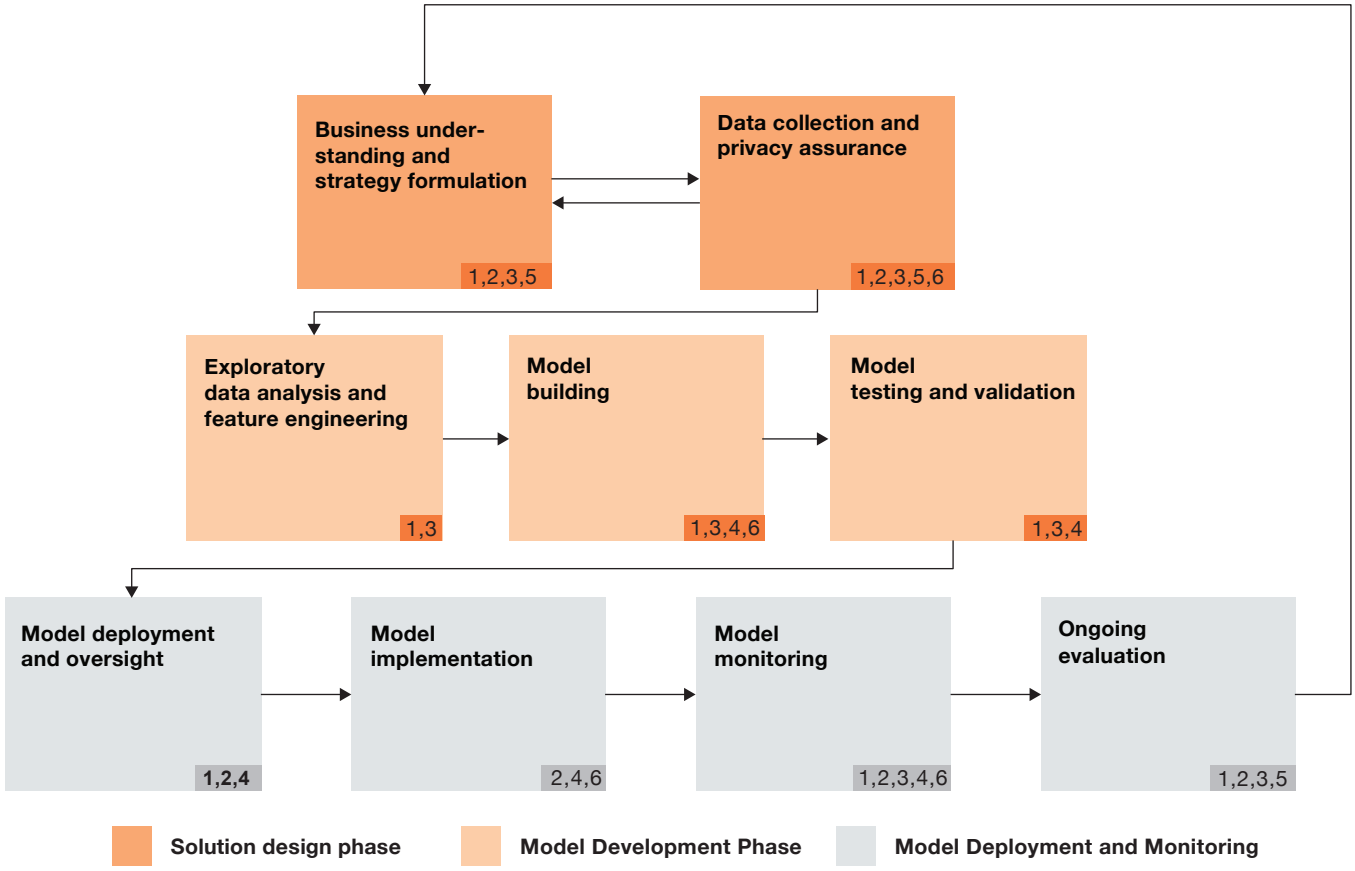
Figure 2 briefly discusses the key gains and risks of using AI in digital lending, mapping them to the phases that they are most likely to be realised in. This helps in identifying the RTAI attributes which are most relevant to each phase. While the primary aim of this section is to emphasise the application of RTAI principles across each phase in the MDLC, the paper also underscores how these principles are equally important to the other AI applications that support the MDLC within digital lending.

**Figure 2:** Benefits and risks of using AI in digital lending and appropriate RTAI attributes to tackle them

| | Gains from AI deployment | Risks from AI deployment | Relevant RTAI Principles |
|---|---|---|---|
| **Solution design and data preparation phase** | Advanced processing capabilities, including processing unstructured data create deeper data to assess credit worthiness | Bias in data, data poisoning may diminish the accuracy of the predictions. Advanced processing may allow for inferences that infringe on the privacy of borrowers. | • Transparency, explainability and contestability<br>• Fairness and non-discrimination<br>• Privacy and data protection<br>• Governance and accountability |
| **Model development phase** | Scalability, adaptability, autonomous learning allowing for credit-decisioning at scale and at potentially lower costs. | The model may run the risk of propagating bias at scale that may keep worthy borrowers from loans or vice versa.<br><br>Data Hallucinations may mislead borrowers diminishing their financial well -being.<br><br>The model may be complex, making it difficult to validate and justify its logic. | • Transparency, explainability and contestability<br>• Fairness and non-discrimination<br>• Privacy and data protection<br>• Dependability of the AI system<br>• Governance and accountability |
| **Model development and monitoring phase** | Efficiency gains, systematic collection of meta data that generates insights about the performance of the model, portfolio quality, gaps in product design etc. | Lack of control over the model, especially if external conditions change, vulnerability to external shocks and manipulation. | • Transparency, explainability and contestability<br>• Fairness and Non-discrimination<br>• Privacy and data protection<br>• Preserving human agency and instituting human oversight<br>• Governance and accountability |

**Figure 3** is a simplified version of the MDLC. This is an iterative process, and the steps may not strictly follow a linear direction. Examining the phases more closely helps in understanding the interlinked aspects of the process and reiterates the need to craft governance frameworks not only for AI outcomes or the product safety approach, but also about how the systems are developed, or the value chain approach, for developing governance frameworks which are aligned with RTAI's principles.

**Figure 3:** MLDC for digital lending




01 Transparency, explainability and contestability

02 Accountability, governance and oversight

03 Fairness and non discrimination

04 Dependability of the AI system and risk mitigation

05 Protecting human agency and instituting human oversight

06 Privacy and data protection

## RTAI's priorities: What lenders must look out for in each phase

To adopt AI responsibly, lenders need to pay close attention to each phase across the entire development lifecycle. This section discusses the three key phases—solution design, model development and model deployment, and monitoring of the lifecycle—and highlights the key aspects lenders should focus on to ensure that the outcomes of AI adoption are aligned to RTAI's principles.

### 1. Solution design

The solution design phase comprises two essential tasks—(i) focusing on the needs of the business, the requirements of the AI system, and (ii) identifying the appropriate model, data sources, and procuring and cleaning the data to make it fit for use for the AI system. In this phase, risks could arise from poor fitment of the model, and not adhering to customer's concerns while selecting the model. Risks may also arise from using biased or obsolete data. Privacy risks may arise if the data is not collected with proper, informed consent or if the collection and processing processes haven't adhered to the principles of data minimisation and purpose limitation.

RTAI practices and tools that could help lenders safeguard themselves against these risks include:

**(i) Establishing ethical review boards and governance frameworks:**

This practice emphasises the need to define roles for model oversight and accountability that are guided by governance frameworks and drafting a policy checklist for the model development process from stakeholders.

**(ii) Addressing the concerns of diverse stakeholders when determining the details of the AI model**

A credit decision typically affects and is affected by a wide range of actors. For instance, a credit decision must necessarily incorporate the financial interests of the lender, the prudence of the regulator and the interests of the customer. However, these interests are not always compatible and can often be in conflict with each other. It is essential that the model prioritises the right metrics. Hypothetically, a model that exhibits high accuracy may bode well for portfolio quality, however when dealing with financially vulnerable customers, even a relatively low error rate may cause deep distress to the customers. Therefore, it is essential that models optimise to reflect the concerns and priorities of all stakeholders for lending as whole to be responsible. Often, inclusivity is included as an afterthought, and the customers' concerns are restricted to the challenges they face in the UI/UX of the product.

For instance, while it's a common practice to just cater to conventional model performance metrics (like accuracy), these metrics aren't broad enough to quantify model performance when dealing with sensitive cases. Hence the need for purposefully integrating inclusivity and stakeholder priorities into the model's design phase.

**(iii) Adhering to data protection principles and assuring privacy**

Data is the mainstay of any AI model. The quality of the data can have severe implications for the functioning of a model. In digital lending, data quality can have implications on the loan quality, customer experience, default and distress, as well as on the privacy of the borrowers and related cohorts. Data protection practices emphasise being mindful of data sources, identifying data integration platforms and opting for privacy-enhancing techniques such as masking the data before feeding it into the system.

**Figure 4:** RTAI recommendations for the solution design phase

RTAI recommendations for the solution design phase which focuses on encompassing (i) focusing on defining business requirements and formulating strategy as well as (ii) preparing the datadata, which is essential for the model, including:

**Ethical review boards and governance frameworks**

This defines roles for model oversight and accountability, guided by governance frameworks and a policy checklist from stakeholders, and is applied to the model development process.

**Data sources**

Data collection strategies are planned before model development, adhering to regulatory policies for ethical handling and minimising demographic bias with stakeholder involvement in identifying sensitive attributes.

**Data integration**

Data integration platforms should be defined in the solutioning phase. The goal is to ingest data in a manner that protects data privacy and is stable in the long run.

**Data quality**

Data quality criteria must be established to ensure that the data being used will provide reliable and fair outputs.

**Data encryption tools**

Before feeding the data to the model development environment appropriate masking/encryption techniques should be defined.

**Data anonymisation services**

Select suitable anonymisation techniques tailored to different features to ensure that the service meets these needs. Masking sensitive attributes helps mitigate bias by preventing the model from correlating demographic groups with outcomes.

## 2. Model development

In this phase the model is trained on the dataset. This entails data collection, processing, analysis and validating the model and RTAI practices which emphasise:

**(i) Selecting appropriate algorithm frameworks**

Consider algorithms that are compatible with the limitations of the data at hand such as imbalanced datasets. Other considerations when selecting an appropriate algorithm includes ensuring that the data is free of any biases and is transparent and accurate.

**(ii) Addressing the scope of bias and the working of the model**

This is done by deploying fairness assessment tools. For example, deploying tools to interpret the model and determine the primary decision-making factors.

**(iii) Identifying parameters of interest and set up systems to track them:**

At this stage, parameters are designed and systems are installed for tracking them. For instance, identifying a suite of proxies for fairness, inclusivity and resilience and setting up dashboards that collect the data on the proxies, parse through them and flag discrepancies/divergences.

**Figure 5:** RTAI recommendations for model development phase

RTAI recommendations for the model development phase, where data is processed and analysed* to build, train and validate the model include:

**Feature store:**

Maintain comprehensive documentation for each feature, establish clear policies for data quality, privacy and ethics, and assign data stewards for governance compliance.

**Algorithm selection framework:**

Consider algorithms that can handle imbalanced datasets and reduce biases. To assure transparency without compromising on accuracy, adding in a proxy/ surrogate interpretable model that mimics the main model's outputs.

**Model training pipelines:**

Incorporate techniques like re-sampling, re-weighting or synthetic data generation in the training pipeline. Use techniques like adversarial debiasing and data masking to further address any potential bias.

**Bias and fairness assessment tools:**

Analyse outcome distributions across demographics using metrics like disparate impact ratio, false rejection rate and statistical parity to detect bias. Implement corrective measures when biases are found.

**Model interpretability tools**

Techniques like shapely additive explanations (SHAP) and local interpretable model-agnostic explanations (LIME) help in understanding the contribution of different features to model predictions.

**Performance monitoring tools**

Create real-time performance dashboards that include model performance and fairness metrics and implement mechanisms to flag any discrepancies.

## 3. Model deployment and monitoring

This phase focuses on integrating the model into the processes, deploying it at full scale and closely monitoring its performance. RTAI's recommendations during this phase emphasise establishing protocols for managing external stakeholders, who may either be third-party service providers essential to the AI system's operation (such as software vendors or cloud service providers) or entities responsible for monitoring and scrutinising the system (e.g. regulators), by:

**(i) Managing external counterparties:**

The model deployment would likely require several parties, external or in-house, to work collectively with the AI systems with different levels of access and responsibilities. Since the data and systems have legal regulations, it is important to design contracts that highlight the roles, responsibilities, liabilities and indemnities of each party to allocate responsibilities.

**(ii) Implementing data management protocols:**

In this phase, the deployers need protocols and processes to take care of data that may no longer be needed.

**Figure 6:** RTAI recommendations for model deployment and monitoring phase

RTAI recommendations for the model deployment and monitoring phase, which focuses on model integration into operation and continuous performance monitoring, include:

**Cloud services or on-premises servers:**

Use containerisation tools to ensure that the development environment can be replicated. Use encryption for data at rest and in transit. Employ secure access controls and regular audits.

**API management systems**

- Implement rate limiting and quotas to prevent abuse and ensure fair usage.
- Monitoring and logging systems.
- Set up alerting systems to notify administrators of any anomalies or deviations from expected behaviour.

**Access management systems**

Implement role-based access control (RBAC) and multi-factor authentication (MFAs) to authenticate and manage who uses the system.

**Regulatory compliance management systems**

Conduct regular compliance audits to ensure ongoing adherence to relevant laws and regulations.

**User support channels**

Create self-service portals for loan information and AI criteria, enabling data modification with logs. Provide processes for individuals to opt out of data collection and to appeal process for AI decision review.

**Data retention/disposal protocols**

Ensure that the data is encrypted and securely deleted using industry-standard practices.

# 3 Implementing RTAI in digital lending

RTAI principles provide a set of practices that can help digital lenders conduct their processes with responsible and trustworthy AI. A checklist of 37 questions that cover the essential components of responsible and trustworthy AI, particularly in the context of digital lending in India has been developed by Dvara Research and PwC to ensure that businesses can adhere to RTAI's principles.

Designed as an interactive, web-based assessment toolkit, the checklist enables the technology teams to take a simple assessment and gauge the strengths of their processes in relation to the six dimensions of the RTAI framework. The web-based assessment toolkit can be accessed **here**. The list of all questions is set out in the Annex I.

## How does the web-based assessment toolkit work?

- The questionnaire allocates different weightage to different practices, reflecting that not all best practices carry equal significance. For instance, questions related to the monitoring and auditing of AI systems are assigned greater weight compared to those addressing disaster management. This distinction arises because, although disaster management remains important, such events occur less frequently in practical, real-world scenarios.
- It also maps the level to which lenders adhere to these practices. For instance, if the ideal cadence of a particular audit is annual, but lenders do it only once every other year, it will get reflected in their scores.

## Intended outcome

- These scores ultimately allow digital lenders to gauge how far their current AI safeguards are from the desired level and how they might close this gap.

# 4 Way forward

Digital lending in India is projected to grow to 500%, reaching a market size of 1.3 trillion USD by 2030.[6] As this growth accelerates, lenders are expected to increasingly explore advanced AI-driven use cases. However, as indicated by the RBI's FREE AI Committee[7], concerns including but not limited to bias and lack of explainability, continue to make lenders hesitant to adopt the full spectrum of AI solutions. This toolkit is designed to help lenders move beyond these constraints by suggesting methods to overcome these reservations and allow for more creative uses of the AI while protecting the interests of the customers.

The toolkit discussed in the paper is aligned with the seven sutras of Responsible AI as recommended by RBI's FREE AI Committee. Each practice in the toolkit presents a way to anchor the AI adoption as per RBI's sutras while being mindful of the constraints faced by legacy systems. Looking ahead, adherence to this toolkit and its adoption by a self-regulatory organisation (SRO) has the potential to meaningfully shape responsible AI use in digital lending. It would not only ensure responsible lending but also operationalise Recommendation 26 of the FREE AI report[7] which calls for an SRO-led toolkit to facilitate the adoption of RTAI by digital lenders and their partners. Designed to be a living tool, the toolkit's long-term relevance will depend on its ability to remain current and responsive to the emerging risks. This is the first iteration of such a checklist for implementing RTAI, and its future versions would benefit from continuous, high-frequency data collection regarding the evolving industry practices and the collective understanding of risks.

6 https://inc42.com/buzz/digital-lending-become-1-3-tn-market-2030-india/.

7 https://rbidocs.rbi.org.in/rdocs/PublicationReport/Pdfs/FREEAIR130820250A24FF2D4578453F824C72ED9F5D5851.PDF

# Annex I

List of questions for the interactive, web-based RTAI self-assessment.

**Transparency, explainability and contestability**

1. To what degree have you incorporated various tools and techniques to ensure that your data collection is done in a manner that is clear and unambiguous to the user? Please state if any consent forms are used or user awareness programmes being conducted to support your rating.
2. To what degree have you employed explainable AI techniques to address the black box nature of the models? To what extent do you put in formal documentation of any and all processes in the MDLC?

**Fairness and non-discrimination**

3. To what degree are there fairness and governance considerations in place in the training data? Mention any steps taken to ensure that the data being used is representative of the entire population demographics
4. To what degree is there a governance policy in place to define the fairness rules in the models? Have you addressed any potential for implicit bias in your model?"
5. To what extent have you employed tests/mechanisms to check for bias in your model's outputs? Have you employed any fairness metrics as part of your model evaluation process? How do you maintain a balance between improving accuracy and addressing bias in cases of conflict?
6. To what extent are there processes to cure any biases in the models?

**Technological dependability of the AI system**

7. How do you keep track of model performance?
8. What kind of controls have you put in place to be able to address any disruptions/ unwanted situations (E.g. disruption could be due to underlying algorithm being unresponsive/server/infra related disruptions)?
9. To what degree are there security processes in place to prevent unintended use of AI?
10. To what degree are there controls in place to disable the algorithm if there is an issue or unintended use of the AI solution is detected?
11. What steps have you taken to prevent your GenAI applications from generating harmful content?
12. How do you mitigate the risk of model hallucination or false information generation in financial advice when interacting with customers?
13. Do you have any mechanisms in place to segregate external untrusted content from user prompts?
14. Do you have any protocols that ensure GenAI applications are thoroughly tested before putting them up in production environment?

15. How prepared is your organisation with regards to tackling any attacks/incidents/threats that may find its way to your systems through AI/Gen AI applications? Have you taken any steps to inform your team of these threats?
16. What mechanisms have you put in place to ensure that your GenAI application is able to identify and tackle adversarial prompts?
17. Do you have any specific benchmarks that are utilised to assess the performance of the LLM in comparison to others in the market?
18. Is there a documented schedule for updating the data used in the model?
19. What measures have been put in place to ensure that the data remains current and representative over time?
20. What strategies are employed to prevent overfitting and how do you ensure that the model effectively generalises unseen data?

**Privacy and data protection**

21. To what degree are there data privacy considerations to protect personally identifiable information (PII)? Do you have different techniques based on the level of sensitivity/ confidentiality that has to be maintained? Is this in-line with the practices suggested by various data governance frameworks?
22. To what degree are there data encryption provisions in place to secure the data? Have you put in considerations for data security in transit?
23. To what extent is there a special level of security access to interact with the production models?
24. To what degree is there a process in place to delete data for those who wish to remove their data from the database?
25. Does your organisation have established data classification protocols to ensure the proper handling and protection of sensitive information?
26. Please indicate the deployment environment for your large language model (LLM).

**Protecting human agency and establishing human oversight**

27. To what degree are there human in the loop (HITL) governance protocols implemented for the model outcomes? Indicative of presence of human underwriters/model stewards.
28. To what extent have you taken into consideration the various guidelines set by governing bodies like RBI and the government (the DPDP Act) to protect your customer's rights?
29. To what degree have you employed feedback systems across the MDLC? This includes both feedback by developers in the process as well as the users.
30. To what measure have you put in facilities that give the user the right to interact with their data? Is the process transparent enough for the user to be able to get a basic understanding of the model outcomes?
31. Have you adopted any measures to involve human oversight over GenAI model interactions to protect your application against malicious attacks (prompt attacks, jailbreaks, etc.)?

**Governance and accountability**

32. To what degree are there capabilities in place to audit models?
33. Do you have clearly defined roles in your organisation pertaining to the different aspects of model life cycle?
34. To what extent have you established measures to ensure that your employees are aware of the pillars of RAI? Have you considered conducting trainings/awareness programmes for the same?
35. To what degree are there capabilities in place to provide visibility to regulators on explainability? Explain if any measures used to describe how your model is able to arrive at an outcome in your statement? Do these meet conditions set by regulators?
36. What is your guide to governing the use of third-party solutions/applications?

# About Dvara Research

Dvara Research is an independent, non-partisan, not-for-profit policy research institution based in India. Its mission is to ensure that every low-income household and every small enterprise has complete access to suitable financial services and social security through a range of channels that enable them to use these services securely and confidently. Since 2008, Dvara Research has deeply analysed, and carefully written about, financial inclusion and social protection in India from policy, regulatory, and practitioner perspectives that are anchored to its mission. Its work has gained the admiration and respect of policymakers and regulators, and since its inception, Dvara Research has been a research-partner of choice for such key policy-making bodies as the Reserve Bank of India, Securities and Exchange Board of India, Pension Fund Regulatory and Development Authority etc.
www.dvararesearch.com

## Contact us

**Indradeep Ghosh**
Executive Director

## Authors

**Beni Chugh, Manvi Khanna**

# About FACE

Fintech Association for Consumer Empowerment (FACE) is the RBI-recognised Self-regulatory Organisation in the FinTech Sector (SRO-FT). FinTech companies of all kinds come together at FACE to build an industry that enables customer-centric financial services that are safe, suitable, and transparent, delivering positive impacts on society and the economy. The FinTech community (providers, enablers, stakeholders, and others) unite at FACE to develop a nurturing ecosystem where companies and consumers responsibly participate and thrive in the digital economy.

https://faceofindia.org/

## Contact us

**Sugandh Saxena**
Chief Executive Officer

## Authors

**Sugandh Saxena**

# About PwC

**We help you build trust so you can boldly reinvent**

At PwC, we help clients build trust and reinvent so they can turn complexity into competitive advantage. We're a tech-forward, people-empowered network with more than 364,000 people in 136 countries and 137 territories. Across audit and assurance, tax and legal, deals and consulting, we help clients build, accelerate, and sustain momentum. Find out more at www.pwc.com.

PwC refers to the PwC network and/or one or more of its member firms, each of which is a separate legal entity. Please see www.pwc.com/ structure for further details.



## Contact us

**Rajnil Malik**
Partner, AI GTM Leader
PwC India

**Vivek Belgavi**
Partner, Financial Services Advisory Leader
PwC India

## Authors

**Neeraj Sibal, Nikita Ann George**

## Editorial

**Rubina Malhotra**

## Design

**Shipra Gupta**

pwc.in



SG/March 2026 - M&C 50808