VIEW POINT

EXPLAINING THE UNEXPLAINABLE IN AI

Why financial services leaders must focus on Explainable AI to achieve their AI ambitions

aws

Infosys®
Navigate your next

Executive Summary

Financial services institutions are accelerating their use of advanced AI systems, but many are now discovering an uncomfortable truth: as AI models become more powerful, they are also becoming harder to understand.

The industry is reaching a point where technological innovation is no longer the limiting factor in AI transformation; explainability is. Boards, regulators, customers, and internal risk teams are all asking the same question: 'Can we trust the decisions these AI systems make?' And increasingly, the answer is 'We just don't know'. Certainly, not without better tools and insights, better governance, and better methods of explanation.

Explainable AI (XAI) refers to an institution's ability to understand, justify, and defend how an AI system reaches a given output. In a highly regulated sector like financial services, built on trust, transparency, and auditability, explainability is fundamental. Yet achieving it is becoming dramatically harder.

What makes this challenge especially pressing is the nature of modern AI itself. The industry has shifted from predictable, rules-based automation to adaptive, autonomous systems that learn continuously, behave non-deterministically, and operate across distributed environments. Deep learning, large language models, and, more recently, agentic AI are transforming how decisions are made, but they also obscure the logic behind those decisions. AI outputs can vary and behaviors can drift. Chains of autonomous actions become harder to trace. All of which introduces a strategic and operational risk that organizations cannot ignore.

The explainability gap is widening precisely at the moment when AI is moving into the industry's most sensitive domains: credit decisioning, fraud detection, trading, customer advice, operational automation, and now multi-agent systems capable of acting independently. Without clear, defensible explanations, institutions face limitations on adoption, increased regulatory exposure, and a growing trust deficit with customers and stakeholders.

This paper examines the explainability challenge in depth, arguing that it is now one of, if not the, biggest barrier to scaling AI safely and sustainably in financial services. It explores:

Why modern AI is inherently harder to explain, and how non-deterministic behaviors, gentic autonomy, and de-centralized AI decision-making are compounding the problem.

The unique implications for financial services, including regulatory compliance, operational risk, and the erosion of trust.

A multi-method approach to Explainable AI, outlining the complementary techniques, from chain-of-thought extraction to Shapley values, required to build a defensible understanding of model behavior.

How institutions can make explainability a strategic priority, embedding it into governance, oversight, and the full AI lifecycle.

The message is clear: AI will not reach its full potential in financial services unless explainability advances in parallel. Institutions that treat XAI as a strategic imperative will be the ones able to scale AI confidently, safely, and at pace.

Introduction

Artificial intelligence has become central to how financial institutions operate, compete, and innovate. From fraud detection to credit decisioning, from automated trading to customer engagement, AI now underpins many of the industry's most critical processes.

AI deployments have moved far beyond simple automation. Today's models are adaptive, self-optimizing, and increasingly agentic, forming chains of intelligence that operate across entire value chains.

Yet as organizations scale AI technologies, one barrier is increasingly acting as a brake on AI transformation and preventing organizations from maximizing the game-changing potential of AI-driven technologies: 'Explainable AI' or 'XAI'.

Explainability refers to a stakeholder's ability to understand how and why an AI system produces a given output. In traditional models, this was straightforward. Rules-based systems produced traceable, auditable logic.

But modern AI systems, especially those based on deep learning, large language models, and reinforcement architectures, are non-deterministic. You put something in, but the output is never quite the same. The model learns. It reacts to new patterns and evolves over time.

This continual improvement and adaptability is the very thing that makes AI powerful. But it also introduces profound uncertainty. It becomes far harder to predict outcomes, and to explain, let alone prove, why an AI system has reached a decision and taken a given action.

Within financial services, a sector built on trust and tight regulatory controls, Explainable AI is rapidly becoming a critical barrier to progress. The industry simply cannot scale AI on a sustainable basis without explainability. But, as many organizations are now finding, explainability is becoming infinitely harder.

The explainability problem: from sausage machine automation to non-deterministic intelligence

For decades, automation technologies operated with reassuring predictability. Traditional IT systems were akin to industrial machinery: a clearly defined set of inputs went in, and the same output reliably came out the other side. These systems were deterministic, and their determinism was the foundation of risk management. Audit trails made sense, root-cause analysis was easy, and explanations were straightforward.

But modern AI systems have changed all of that. Today's models, particularly machine learning and deep learning architectures, exhibit non-deterministic behavior by design. You can repeatedly provide the same input but the outputs you get will vary slightly from one request to the next. This is because AI models adapt to new data, refine their internal states, and react dynamically to real-world signals. Their behavior evolves over time, and this is precisely what makes them so powerful.

But with that evolution comes uncertainty. And uncertainty is never welcome in highly regulated industries such as financial services and healthcare.

When an AI model declines a loan application, adjusts a trading strategy, flags a transaction as anomalous, or recommends a course of action to a customer or employee, institutions must be able to explain the rationale. They must demonstrate consistency, fairness, and a clear chain of logic. In a deterministic system, this is easy. In a non-deterministic one, explanations become elusive. If the outcome is questioned, by a regulator, a customer, or an internal audit team, the organization needs to be able to justify the decision. But that's becoming increasingly difficult.

This is the essence of the explainability crisis. And more than technology innovation or data quality, it is now becoming the biggest barrier to the proliferation of AI solutions across the financial services industry. Without explainability, there can be no trust. And without trust, AI transformation will never reach its potential.

The Perfect Storm: Why Explainability is Becoming Harder

The challenge of explainability is becoming exponentially more difficult due to three major trends that are converging at the same time.

Firstly, there is ever-greater complexity within AI models. The internal structure of today's AI has become so intricate that even model creators cannot fully describe how individual inferences are produced. Layers of representation form spontaneously and relationships emerge between billions of parameters without any human intervention. Inference behavior may depend on statistical patterns buried deep within training data that no one has examined in years.

This does not mean the models are untrustworthy. It means they are fundamentally different from traditional systems. They are not machines following predefined rules but complex, adaptive systems with behaviors that arise from the interplay of data, training dynamics, and emergent internal representations. As these architectures grow, the distance between input and explanation widens. And yet they are being deployed in high-stakes contexts where clarity is essential.

Secondly, there is the rapid rise of agentic AI and the compounded drift this causes. The arrival of agentic AI is adding an entirely new layer of complexity. Instead of a single AI model making a single prediction, organizations are deploying networks of intelligent agents capable of taking actions, invoking tools, collaborating, and responding to real-time environments. In these systems, outputs are no longer static predictions but chains of autonomous decisions.

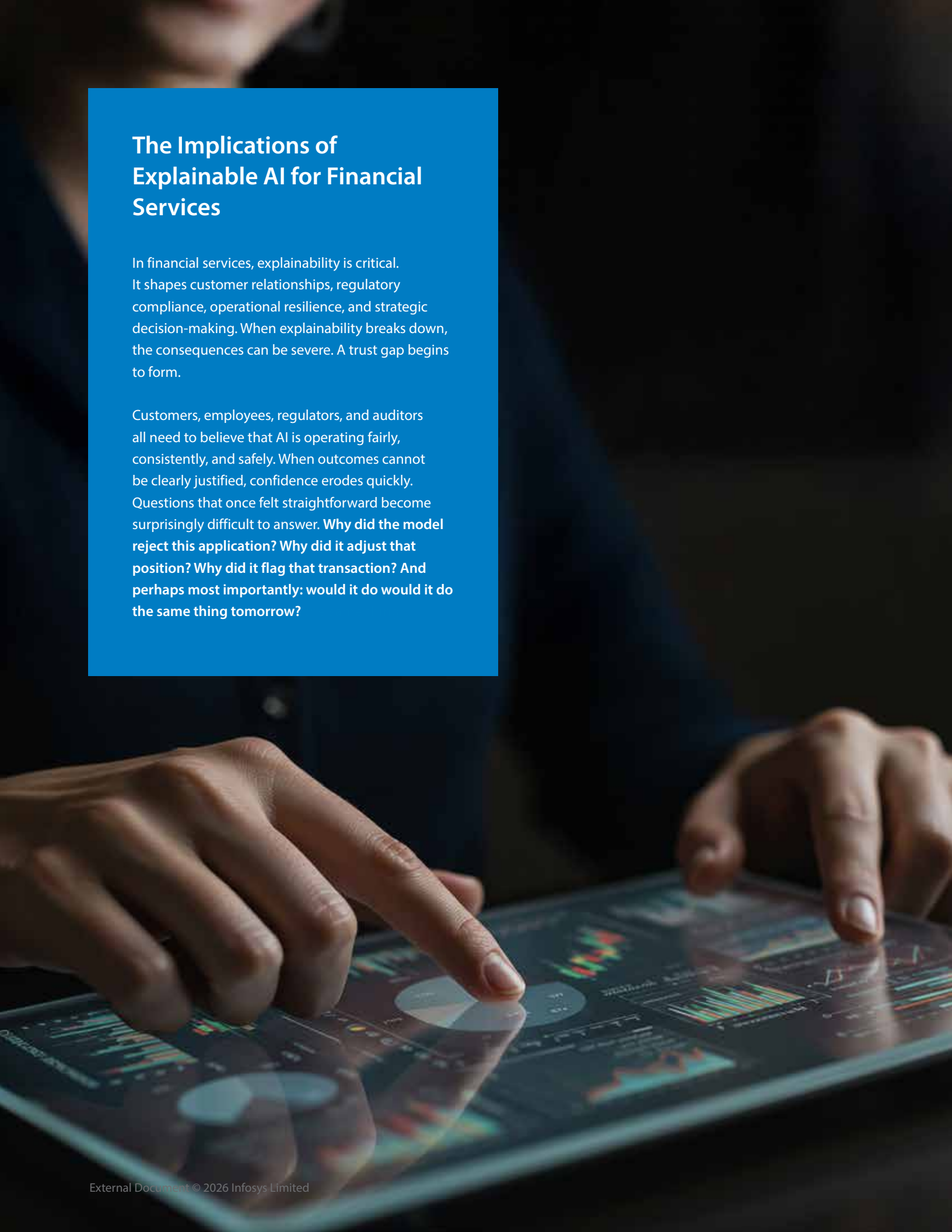
With this autonomy comes drift, not just within individual models, but drift that compounds across the entire agentic value chain. A small degree of variability within one agent's behavior may be insignificant on its own. But when a dozen interconnected agents are each introducing a small percentage of error or deviation, the cumulative effect becomes far more substantial. And the end result may be materially different from what any one agent intended. Explaining the behavior of one AI model is difficult.

Explaining the behavior of ten interacting models acting autonomously in real-time is something else entirely. This compounded drift is one of the primary reasons regulated industries are approaching agentic AI with both excitement and caution. The potential value is undoubtedly immense but the explainability challenge is even greater.

Finally, the third trend which is exacerbating the problem of explainability is the shift to decentralized, edge-driven decision-making. Financial services institutions are no longer tightly controlled hierarchical systems where decisions are taken within a central core. Increasingly, intelligence is being pushed outwards, into branches, trading gateways, mobile devices, fraud sensors, and countless other points at the edge of the organizational network.

This decentralization improves performance, resilience, and customer experience. But it also fragments decision-making. Each node may operate with slightly different data, environmental conditions, or model versions. Behavior begins to diverge across locations and oversight becomes more complex. Explanations become harder to assemble because the full context of an AI-driven decision may only exist at the edge.

Together, these three forces - AI model complexity, agentic AI autonomy, and decentralized decisioning on the edge - have created a perfect storm. Explainability has not simply become more important; it has become more elusive.



The Implications of Explainable AI for Financial Services

In financial services, explainability is critical. It shapes customer relationships, regulatory compliance, operational resilience, and strategic decision-making. When explainability breaks down, the consequences can be severe. A trust gap begins to form.

Customers, employees, regulators, and auditors all need to believe that AI is operating fairly, consistently, and safely. When outcomes cannot be clearly justified, confidence erodes quickly. Questions that once felt straightforward become surprisingly difficult to answer. **Why did the model reject this application? Why did it adjust that position? Why did it flag that transaction? And perhaps most importantly: would it do would it do the same thing tomorrow?**

Without clear answers, trust in the system falters and risk exposure increases, often in unpredictable ways.

Non-deterministic systems introduce new categories of risk that traditional governance structures were not designed to handle. Model drift becomes a constant threat; bias can emerge unintentionally through subtle shifts in data distributions; hallucinations appear unpredictably in generative models; and inconsistencies accumulate.

In fast-moving environments, particularly trading and fraud detection, small instabilities can escalate quickly. The industry has witnessed this dynamic, long before modern AI entered the picture.

Indeed, 'Ghost in the machine' disasters are nothing new. The collapse of Knight Capital, the largest trader in US equities in 2012, remains one of the most striking examples. A forgotten piece of code, left dormant in a DevOps pipeline, was inadvertently triggered in a live trading environment. Over the course of a single day, the firm lost hundreds of millions of dollars. There were no effective circuit breakers, no human oversight fast enough to intervene, and no clear understanding of what the system was doing.

Significantly, this was not even an autonomous AI model; it was automation that had become too complex and too fast for its human overseers to manage. But the parallels with modern AI are impossible to ignore.

As systems become more intelligent and more autonomous, the potential for 'ghost in the machine' scenarios is amplified. Without explainability, risk cannot be fully understood or effectively mitigated.

In response, governments and industry authorities are moving to ensure that organizations are taking the challenge of explainability seriously.

The EU AI Act represents the most comprehensive attempt yet to regulate Artificial Intelligence technology deployment. Financial services use cases fall overwhelmingly into the "high-risk" category as defined by the legislation, bringing with them stringent obligations and significant fines - up to 7% of global turnover - for non-compliance.

Yet the legislation defines risk in broad conceptual terms (such as bias, hallucination, and drift) without offering concrete thresholds. Institutions are left to interpret expectations on their own. The result is a landscape where organizations know they must comply but are uncertain how to do so, slowing the adoption of AI, and increasing anxiety at the board and executive levels.

In short, explainability is not simply a technical challenge; it is a strategic and regulatory imperative. And leaders urgently need to find answers.

Tackling XAI With a Multi-Method Approach

The instinctive reaction to a challenge of this scale is to look for a definitive fix - a new tool, method, or governance model that can remove all uncertainty. But unfortunately, intelligence, whether human or artificial, is not that tidy. As people, we struggle to provide perfect explanations for our decisions. Certainty is not our natural state, and it is not AI's either.

What organizations need instead, is a portfolio of approaches and a set of complementary methods that together start to provide a reasonable, defensible view into how AI systems behave. There are dozens of such techniques in the field today. None is sufficient on its own but together they form the beginnings of a solution to this pressing challenge.

One increasingly common method is chain-of-thought extraction, which interrogates the AI model and encourages it to articulate the steps it took to reach a decision. This does not always reveal a perfect internal map, but it offers clues about the logic and assumptions being used, and it highlights inconsistencies that might otherwise go undetected.

Another widely used approach is LIME (Local Interpretable Model-Agnostic Explanations), which builds a simplified model that approximates the original AI's behavior for a specific prediction. Instead of revealing the entire inner workings of a complex architecture, LIME provides a localized, human-readable analogue that shows why the model behaved as it did in a particular case. In regulated environments, this can be invaluable.

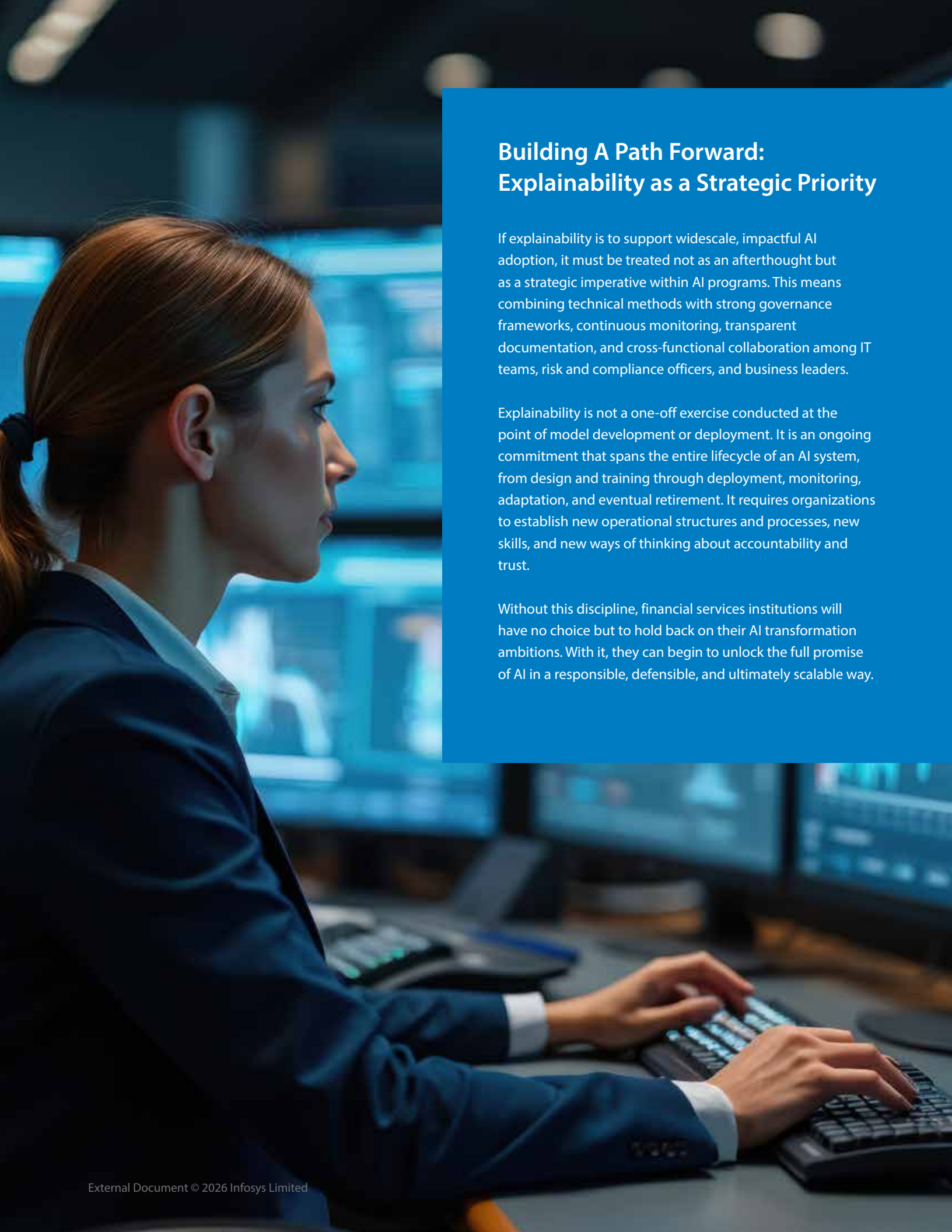
Counterfactual analysis is also gaining momentum because it reframes explainability around the concept of change. By identifying what adjustments would be required to reach a different outcome, organizations can understand which variables truly matter, how sensitive the model is to shifting conditions, and whether unfair or biased influences may be present.

Finally, Shapley methods offer one of the most mathematically robust ways to attribute influence across features. By assessing how each input contributes to the output across all possible combinations, Shapley values provide a global, quantitative view of model behavior that aligns well with regulatory expectations and risk frameworks.

Already there are 20 or 30 different ways to analyze how an AI system behaves and each technique tells a different part of the story. But together, they create a richer, more reliable picture of the system's behavior than any single method could achieve. Leaders need to use a combination of these methods over their AI models to get a consensus view on how the AI is performing and get some degree of certainty as to whether it is working well or not.

Organizations should also be leaning on the expertise and experience of trusted partners. Such is the speed at which this challenge is evolving, leaders from within risk, compliance and IT, need to come together with their technology partners to develop holistic strategies for XAI, taking into account regulatory requirements, soaring levels of complexity, and the ongoing shift towards de-centralized architectures.

Leaders should be looking to their partners to bring a structured perspective on which XAI techniques to combine, how to implement them across the AI model lifecycle, and how to create a unified narrative that risk teams, auditors, and supervisors can understand and get behind. Most importantly, partners should be helping leaders see beyond the technical mechanics of XAI to consider the strategic implications, so that explainability is prioritized and supports broader AI ambitions.



Building A Path Forward: Explainability as a Strategic Priority

If explainability is to support widescale, impactful AI adoption, it must be treated not as an afterthought but as a strategic imperative within AI programs. This means combining technical methods with strong governance frameworks, continuous monitoring, transparent documentation, and cross-functional collaboration among IT teams, risk and compliance officers, and business leaders.

Explainability is not a one-off exercise conducted at the point of model development or deployment. It is an ongoing commitment that spans the entire lifecycle of an AI system, from design and training through deployment, monitoring, adaptation, and eventual retirement. It requires organizations to establish new operational structures and processes, new skills, and new ways of thinking about accountability and trust.

Without this discipline, financial services institutions will have no choice but to hold back on their AI transformation ambitions. With it, they can begin to unlock the full promise of AI in a responsible, defensible, and ultimately scalable way.

Tackling Explainability With The Five Ws Framework

As financial services institutions scale their use of agentic AI, they need more than a collection of technical tools. They need a coherent framework that ensures explainability is addressed comprehensively, consistently, and in a manner that satisfies the diverse needs of regulators, customers, auditors, and internal stakeholders.

The Five Ws of Explainable AI provides such a framework. Rooted in the fundamental questions that any stakeholder might ask about an AI system, it offers a structured approach to building, deploying, and governing agentic AI in a way that is transparent, defensible, and fit for the demands of a highly regulated industry.

The framework comprises five interconnected dimensions, each addressing a critical question that institutions must be able to answer:

What did the agent do?

This is the foundation of explainability: complete visibility into every action an agent takes. In agentic systems, where autonomous agents may invoke tools, access data, interact with other agents, and take consequential actions, institutions must maintain a comprehensive and tamper-evident record of activity. This dimension encompasses not just logging individual actions, but tracing entire sessions, tracking data lineage, and maintaining version histories that enable full reconstruction of agent behavior. Without this transparency, everything else becomes impossible. You cannot explain what you cannot see. And you cannot govern what you cannot observe.

Why did it decide this?

Transparency alone is insufficient. Stakeholders need to understand the reasoning behind agent decisions. When a customer asks why their mortgage application was declined, or a regulator questions a trading decision, the institution must be able to articulate the factors, evidence, and logic that led to the outcome. This requires capturing not just what the agent did, but the chain of reasoning that informed its actions. It means identifying which inputs most influenced the output, grounding responses in verifiable source material, and being able to explain what would need to change for a different outcome.

In an industry where decisions profoundly affect people's financial lives, the ability to provide clear, honest reasoning is not optional.

Who is responsible?

AI does not operate in a vacuum. Behind every agent is a web of human accountability: the teams who designed it, the leaders who approved its deployment, the operators who monitor its performance. This dimension demands clarity on who owns each agent, who is accountable when things go wrong, and where human oversight intersects with machine autonomy. It encompasses defining which decisions require human approval, establishing escalation triggers for unusual situations, enabling humans to intervene and override agent actions, and ensuring that incidents can be traced back to responsible parties. In financial services, where personal accountability is a regulatory expectation and a cornerstone of trust, this dimension is non-negotiable.

What boundaries exist?

Explainability is not only about understanding what happened; it is about preventing what should never happen. This dimension focuses on the constraints that limit agent autonomy and the controls that ensure agents operate within acceptable parameters. What actions are prohibited? What thresholds trigger intervention? What safeguards prevent an agent from accessing data it should not see or taking actions beyond its authority? These boundaries must be explicit, enforceable, and auditable. But this dimension also encompasses ongoing evaluation of agent performance: monitoring for accuracy, consistency, fairness, and drift over time. Boundaries are not static; they must be continuously validated against real-world behavior. An agent that operated safely yesterday may not operate safely tomorrow if its behavior has shifted. Institutions must therefore combine preventive constraints with continuous quality monitoring to ensure that agents remain within acceptable bounds throughout their operational life.

What governance is in place?

The final dimension brings everything together. Governance provides the organizational structures, processes, and controls that ensure explainability is maintained throughout the AI lifecycle, from initial design through deployment, operation, and eventual retirement. It encompasses policy management and version control, documentation standards and model cards, roles and responsibilities across the three lines of defense, audit and assurance processes, incident management procedures, and continuous improvement mechanisms. Without governance, the other dimensions remain fragmented and unsustainable. With it, institutions can demonstrate to regulators, boards, and customers that explainability is not an afterthought but an embedded discipline.

Together, these five dimensions form a comprehensive approach to explainability that is specifically designed for the challenges of agentic AI in financial services. The framework recognizes that different stakeholders need different types of explanations. A regulator examining systemic risk needs to see boundaries, governance structures, and audit trails. A customer questioning a decision needs to understand the reasoning in plain terms. An internal auditor needs assurance that controls are operating effectively and that accountability is clearly assigned. The Five Ws framework ensures that institutions can satisfy all of these needs in a coherent and consistent manner.

Critically, the framework is not merely theoretical. Each dimension maps directly to capabilities that can be implemented today using a combination of technology, process, and organizational design. The question for financial services leaders is not whether to adopt such a framework, but how quickly they can embed it into their AI transformation programs.

Partnering With Infosys and AWS

Implementing the Five Ws framework requires a combination of deep technical capability, industry expertise, and strategic vision. Infosys and AWS have joined forces to provide financial services institutions with an end-to-end solution that makes this framework operational.

AWS provides the foundational infrastructure for explainable agentic AI. Amazon Bedrock AgentCore, now generally available, offers the most advanced platform for building, deploying, and operating AI agents at enterprise scale. Its capabilities directly address each dimension of the Five Ws framework.

For transparency into what agents do, AgentCore Observability delivers complete visibility into agent behavior, creating comprehensive audit trails that capture every action, tool invocation, and agent interaction. Combined with AWS CloudTrail and Amazon CloudWatch, institutions gain unified logging and monitoring across their entire AI estate, regardless of where agents operate.

For understanding why agents decide as they do, Amazon SageMaker Clarify provides sophisticated techniques including Shapley value analysis for feature attribution and counterfactual explanations that reveal what would need to change for a different outcome. Amazon Bedrock Guardrails adds grounding capabilities that link agent responses to verified source material, ensuring that explanations are rooted in evidence rather than fabrication.

For establishing who is responsible, AgentCore integrates with AWS Identity and Access Management to provide clear ownership and access controls. AgentCore Policy enables institutions to define human-in-the-loop requirements for high-stakes decisions, while CloudWatch Alarms create escalation triggers that bring human judgment into the loop when conditions warrant.

For defining what boundaries exist, AgentCore Policy provides precise, auditable constraints using Cedar, an open-source policy language that is both machine-executable and human-readable. Policies can incorporate authentication claims, transaction values, and business context to enforce fine-grained permissions. AgentCore Evaluations complements these preventive controls with continuous quality monitoring, measuring dimensions such as correctness, helpfulness, fairness, and safety against defined thresholds. When agent behavior drifts or quality degrades, the system alerts operators before small issues compound into significant problems.

For ensuring what governance is in place, AWS Audit Manager automates evidence collection against compliance frameworks, while infrastructure-as-code practices enable version-controlled policy management. AWS AI Service Cards provide standardized documentation templates, and the integration of all these services into a unified console enables coherent oversight across the AI lifecycle.

Infosys brings the strategic and implementation expertise required to translate these technical capabilities into business value. The Infosys Responsible AI toolkit provides a suite of APIs that integrate safety, security, privacy, explainability, fairness, and hallucination detection directly into AI solutions. The toolkit's Explainability module supports both traditional machine learning and deep learning models, offering local and global explanation methods that help institutions answer the "why" questions that customers and regulators ask. The Fairness and Bias capabilities enable institutions to analyze AI responses for potential biases across demographic, socioeconomic, cultural, and geographic dimensions, ensuring that agents treat all customers fairly and consistently.

Beyond technology, Infosys provides the consulting expertise to help institutions design and implement governance frameworks tailored to their specific regulatory environment and risk appetite. This includes mapping the Five Ws framework to regulatory requirements such as the EU AI Act, developing organizational structures and processes for AI oversight, defining roles and responsibilities across business, technology, and risk functions, and building the internal capabilities required to sustain explainability as AI deployments scale.

The partnership between Infosys and AWS reflects a shared conviction that explainability is not a barrier to AI adoption but an enabler of it. Institutions that embed the Five Ws framework into their AI programs will be able to deploy agentic AI with confidence, knowing that they can answer the questions that regulators, customers, and boards will inevitably ask. They will be able to scale AI faster, because trust accelerates adoption. And they will be better positioned to capture the transformative potential of agentic AI while managing the risks that come with autonomous, adaptive systems.

The technology exists. The framework is clear. The path forward is defined. What remains is the commitment to make explainability a strategic priority, and the partnership to make it real.

Author

Vijay Rathore,
Head of AI, Cloud Strategy and Sales, EMEA

Dr Salman Taherian
Global Agentic AI Partner Lead, AWS

For more information, contact askus@infosys.com

Infosys®
Navigate your next