CENTRE DE DROIT
BANCAIRE ET FINANCIER

Explainability of an AI system

Clarifications from the Financial Stability Institute

Par Philippe Gilliéron le 24 September 2025

On 8 September 2025, the Financial Stability Institute (FSI) published a document aimed at clarifying what is meant by ‘explainability of artificial intelligence systems’ and certain options for how to meet this requirement.

This document is particularly important given that FINMA has made explainability one of the key points for banks to consider in its Communication 08/2024 of 18 December 2024 on the supervision of governance and risk management in relation to the use of artificial intelligence (see Caballero Cuevas, cdbf.ch/1392).

It should be noted that ‘explainability’ refers to the ability to explain how the system works and why it has produced the result generated. An essential component of the responsible implementation of AI systems and a decisive factor in gaining customer trust and meeting FINMA’s supervisory expectations, explainability is nevertheless hampered by the complexity of these systems, which are often likened to black boxes.

In many respects, the most powerful AI models (deep neural networks, LLMs) are the least explainable. This is particularly true when banks base their systems on proprietary LLMs, which are the most commonly used. However, existing explainability techniques (SHAP, LIME, counterfactual explanations) have limitations : instability, potential inaccuracy, and lack of a reference truth to assess their relevance.

This situation creates a regulatory paradox. Current guidelines on model risk management (MRM) adopted in various countries implicitly require explainability through governance, documentation and validation requirements. However, when applied strictly, these rules effectively prohibit the most advanced AI models, depriving banks of potentially superior risk management tools.

A. Towards a differentiated and pragmatic approach

Recognising that a binary approach (explainable/unexplainable) is not realistic, the FSI proposes several adjustments to MRM based on the following criteria :

Risk-based categorisation. The required explainability must be proportionate to the criticality of the use case and the complexity of the model. Thus, AI systems limited to internal use, such as chatbots, do not require the same level of transparency in their operation as systems that have a direct impact on customer relations and decision-making or that offer an interface with customers.

Recognition of explainability-performance trade-offs. The FSI considers that the use of systems for critical use cases cannot be ruled out outright due to a lack of explainability, provided, however, that the gap between expected explainability and possible explainability is not too great, a notion that is admittedly subject to caution. In such cases, it is necessary to examine the extent to which the risks resulting from this lack of explainability can be minimised by adopting other measures (increased monitoring, enhanced data governance, human supervision, automatic shutdown mechanisms) that make the residual risk acceptable in view of the level of performance and efficiency of the system concerned, which must, however, appear to be significantly higher than that of simpler and more explainable models.

Systems used for regulatory compliance. The FSI recognises that, for calculation purposes, it may be more complicated to allow the use of tools designed to calculate capital requirements and other regulatory requirements in the absence of explainability. For the FSI, a compromise could be to authorise the use of complex AI models recognised on the market as being effective in these areas for certain categories of risk only, for a certain level of risk exposure, or to provide for risk weighting calculated using these systems with thresholds that are stricter than those provided for in Basel III.

B. Practical recommendations for banking institutions

Without going into detail here about establishing a comprehensive AI tool management system based on standards such as the NIST AI RMF or ISO 42001, the following points constitute measures for implementing the approach proposed by the report :

- Map all AI use cases and classify them according to their level of criticality and explainability.
- Adopt an internal policy defining explainability thresholds appropriate to the risk categories.
- Implement an independent review including SHAP/LIME tests and regular validation throughout the life cycle of these systems.
- Train compliance, risk and IT teams on FINMA requirements and the limitations of the techniques that may be used.
- Provide for measures to minimise the risks resulting from poor explainability.
- Organise a simulated internal FINMA compliance audit on explainability and AI governance.
- Document the bank’s approach to explainability, using concrete use cases demonstrating the risk/customer benefits of complex systems, with justification for any exemptions granted for certain systems and measures taken to minimise risks to an acceptable level.

C. Conclusion

The approach proposed by the FSI offers a possible framework for banking institutions to determine the risks associated with the deployment of AI systems in terms of their level of explainability, assess them and make ‘responsible’ decisions in this area. The challenge is therefore no longer to choose between performance and explainability, but to build a governance framework that allows complex models to be used with an adequate level of control. This change requires a significant investment in skills and processes, but it could considerably accelerate the adoption of AI in critical banking activities. Institutions that anticipate these changes will gain a decisive competitive advantage in the digital transformation of the sector.