CENTRE DE DROIT
BANCAIRE ET FINANCIER

Explicabilité d'un système IA

Les clarifications du *Financial Stability Institute*

Par Philippe Gilliéron le 24 septembre 2025

Le 8 septembre 2025, le *Financial Stability Institute* (FSI) a publié un document visant à clarifier ce qu'il faut entendre par « explicabilité des systèmes d'intelligence artificielle » et certaines options quant à la manière de répondre à cette exigence.

Ce document présente un intérêt d'autant plus important que la FINMA a fait de l'explicabilité l'un des points majeurs devant retenir l'attention des établissements bancaires dans le cadre de sa communication 08/2024 du 18 décembre 2024 sur la surveillance relative à la gouvernance et la gestion des risques en lien avec l'utilisation de l'intelligence artificielle (cf. Caballero Cuevas, cdbf.ch/1392).

Rappelons que, par « explicabilité », on entend la capacité à pouvoir expliquer le fonctionnement du système et la raison pour laquelle il a abouti au résultat généré. Composante essentielle d'une mise en œuvre responsable des systèmes IA et facteur décisif pour gagner la confiance des clients et répondre aux attentes en matière de surveillance de la part de la FINMA, cette explicabilité se heurte toutefois à la complexité de ces systèmes, bien souvent apparentés à des boîtes noires.

À bien des égards, les modèles IA les plus performants (réseaux de neurones profonds, LLM) sont les moins explicables. Cela est d'autant plus vrai lorsque les établissements bancaires font reposer leurs systèmes sur des LLM propriétaires comme le sont les plus couramment utilisés. Or, les techniques d'explicabilité existantes (SHAP, LIME, explications contrefactuelles) présentent des limites : instabilité, inexactitude potentielle, et absence de vérité de référence pour évaluer leur pertinence.

Cette situation crée un paradoxe réglementaire. Les lignes directrices actuelles sur la gestion des risques modèles (*Model Risk Management,* MRM) adoptées en différents pays exigent implicitement l'explicabilité à travers les exigences de gouvernance, documentation et validation. Appliquées strictement, ces règles reviennent toutefois à proscrire les modèles IA les plus avancés, privant les banques d'outils potentiellement supérieurs pour la gestion des risques.

A. Vers une approche différenciée et pragmatique

Reconnaissant qu'une approche binaire (explicable/non explicable) n'est pas réaliste, le FSI propose plusieurs ajustements aux MRM à l'aune des critères suivants :

Catégorisation basée sur les risques. L'explicabilité requise doit être proportionnée à la criticité du cas d'usage et à la complexité du modèle. Ainsi les systèmes IA limités à un usage interne de type *chatbot* n'exigent-ils pas le même niveau de transparence quant à leur fonctionnement que des systèmes ayant un impact direct dans la relation client et la prise de décision ou proposant une interface avec les clients.

Reconnaissance des compromis explicabilité-performance. Le FSI considère que le recours à des systèmes pour des cas d'usage critiques ne saurait être d'emblée exclu en raison d'un manque d'explicabilité, pour autant, cependant, que l'écart entre l'explicabilité attendue et l'explicabilité possible ne soit pas trop important, notion il est vrai sujette à caution. En ces hypothèses, il convient d'examiner dans quelle mesure les risques résultant de ce manque d'explicabilité peuvent être minimisés par l'adoption d'autres mesures (surveillance accrue, gouvernance des données renforcée, supervision humaine, mécanismes d'arrêt automatique) qui rendent le risque résiduel acceptable eu égard au niveau de performance et efficience du système concerné, qui doit toutefois apparaître bien supérieur à celui de modèles plus simples et davantage explicables.

Systèmes utilisés en matière de conformité règlementaire. Le FSI reconnaît que pour le calcul il peut être plus compliqué d'admettre le recours à des outils destinés à calculer les fonds propres nécessaires et autres exigences réglementaires en cas de manque d'explicabilité. Pour le FSI, un compromis pourrait consister à autoriser l'utilisation de modèles d'IA complexes reconnus sur le marché comme étant performants en ces domaines pour certaines catégories de risques uniquement, pour un certain niveau d'exposition au risque ou encore de prévoir une pondération des risques calculée à l'aide de ces systèmes avec des seuils plus stricts que ceux prévus dans Bâle III.

B. Recommandations pratiques pour les établissements bancaires

Sans rentrer ici dans les détails visant à l'établissement d'un système de gestion complet des outils IA à l'aune de standards comme le NIST AI RMF ou la norme ISO 42001, les points suivants constituent des mesures propres à mettre en œuvre l'approche proposée par le rapport :

- Cartographier tous les cas d'usage IA et les classer selon leur niveau de criticité et explicabilité.
- Adopter une politique interne définissant des seuils d'explicabilité adaptés aux catégories de risque.
- Mettre en place une revue indépendante incluant tests SHAP/LIME et validation régulière tout au long du cycle de vie de ces systèmes.
- Former les équipes conformité, risques et IT sur les exigences FINMA et les limites des techniques susceptibles d'être utilisées.
- Prévoir des mesures visant à minimiser les risques résultant d'une explicabilité déficiente.
- Organiser un audit interne simulé de conformité FINMA sur l'explicabilité et la gouvernance IA.
- Documenter l'approche de la banque en matière d'explicabilité, en prenant des cas d'usage concrets démontrant les bénéfices risques/clients des systèmes complexes, avec une justification concernant les dérogations accordées pour certains systèmes et mesures prises pour minimiser les risques à un niveau jugé acceptable.

C. Conclusion

L’approche proposée par le FSI offre une grille de lecture possible pour permettre aux établissements bancaires de déterminer les risques liés au déploiement de systèmes IA eu égard à leur niveau d’explicabilité, les apprécier et prendre des décisions « responsables » en la matière. L’enjeu n’est ainsi plus de choisir entre performance et explicabilité, mais de construire un cadre de gouvernance permettant d’exploiter les modèles complexes avec un niveau de contrôle adéquat. Cette évolution nécessite un investissement significatif en compétences et processus, mais elle pourrait considérablement accélérer l’adoption de l’IA dans les activités bancaires critiques. Les établissements qui anticipent ces évolutions prendront une avance concurrentielle décisive dans la transformation numérique du secteur.
