CENTRE DE DROIT
BANCAIRE ET FINANCIER

Erklärbarkeit eines KI-Systems

Erläuterungen des *Financial Stability Institute*

Par Philippe Gilliéron le 24 September 2025

Am 8. September 2025 veröffentlichte das *Financial Stability Institute* (FSI) ein Dokument, um zu klären, was unter „Erklärbarkeit von Systemen der künstlichen Intelligenz" zu verstehen ist, und um einige Optionen aufzuzeigen, wie diese Anforderung erfüllt werden kann.

Dieses Dokument ist umso interessanter, als die FINMA in ihrer Mitteilung 08/2024 vom 18. Dezember 2024 zur Aufsicht über die Governance und das Risikomanagement im Zusammenhang mit dem Einsatz künstlicher Intelligenz (vgl. Caballero Cuevas, cdbf.ch/1392).

Zur Erinnerung : Unter „Erklärbarkeit" versteht man die Fähigkeit, die Funktionsweise des Systems und die Gründe für das erzielte Ergebnis zu erklären. Diese Erklärbarkeit ist ein wesentlicher Bestandteil einer verantwortungsvollen Umsetzung von KI-Systemen und ein entscheidender Faktor, um das Vertrauen der Kunden zu gewinnen und die Erwartungen der FINMA in Bezug auf die Aufsicht zu erfüllen. Sie stößt jedoch auf die Komplexität dieser Systeme, die oft mit Black Boxes verglichen werden.

In vielerlei Hinsicht sind die leistungsfähigsten KI-Modelle (tiefe neuronale Netze, LLM) am wenigsten erklärbar. Dies gilt umso mehr, wenn Banken ihre Systeme auf proprietäre LLM stützen, wie sie am häufigsten verwendet werden. Die bestehenden Techniken zur Erklärbarkeit (SHAP, LIME, kontrafaktische Erklärungen) weisen jedoch Grenzen auf : Instabilität, potenzielle Ungenauigkeit und das Fehlen einer Referenzwahrheit zur Bewertung ihrer Relevanz.

Diese Situation führt zu einem regulatorischen Paradoxon. Die aktuellen Leitlinien zum Modellrisikomanagement (*Model Risk Management,* MRM), die in verschiedenen Ländern verabschiedet wurden, verlangen implizit Erklärbarkeit durch Anforderungen an Governance, Dokumentation und Validierung. Bei strikter Anwendung führen diese Regeln jedoch dazu, dass die fortschrittlichsten KI-Modelle verboten werden, wodurch den Banken potenziell überlegene Instrumente für das Risikomanagement vorenthalten werden.

A. Hin zu einem differenzierten und pragmatischen Ansatz

In der Erkenntnis, dass ein binärer Ansatz (erklärbar/nicht erklärbar) nicht realistisch ist, schlägt das FSI mehrere Anpassungen des MRM vor, die sich an folgenden Kriterien orientieren :

Risikobasierte Kategorisierung. Die erforderliche Erklärbarkeit muss in einem angemessenen Verhältnis zur Kritikalität des Anwendungsfalls und zur Komplexität des Modells stehen. So

erfordern KI-Systeme, die auf den internen Gebrauch wie beispielsweise *Chatbots* beschränkt sind, nicht das gleiche Maß an Transparenz hinsichtlich ihrer Funktionsweise wie Systeme, die einen direkten Einfluss auf die Kundenbeziehung und die Entscheidungsfindung haben oder eine Schnittstelle zu den Kunden bieten.

Anerkennung von Kompromissen zwischen Erklärbarkeit und Leistung. Der FSI ist der Ansicht, dass der Einsatz von Systemen für kritische Anwendungsfälle nicht von vornherein aufgrund mangelnder Erklärbarkeit ausgeschlossen werden kann, sofern die Diskrepanz zwischen der erwarteten und der möglichen Erklärbarkeit nicht zu groß ist, was allerdings ein mit Vorsicht zu behandelnder Begriff ist. In diesen Fällen sollte geprüft werden, inwieweit die aus dieser mangelnden Erklärbarkeit resultierenden Risiken durch andere Maßnahmen (verstärkte Überwachung, verbesserte Datenverwaltung, menschliche Aufsicht, automatische Abschaltmechanismen) minimiert werden können, die das Restrisiko angesichts der Leistungsfähigkeit und Effizienz des betreffenden Systems akzeptabel machen, wobei dieses jedoch deutlich über dem von einfacheren und besser erklärbaren Modellen liegen muss.

Systeme für die Einhaltung gesetzlicher Vorschriften. Der FSI räumt ein, dass es für die Berechnung schwieriger sein kann, den Einsatz von Instrumenten zur Berechnung der erforderlichen Eigenmittel und anderer regulatorischer Anforderungen zuzulassen, wenn die Erklärbarkeit fehlt. Für den FSI könnte ein Kompromiss darin bestehen, die Verwendung komplexer KI-Modelle, die auf dem Markt als leistungsfähig in diesen Bereichen anerkannt sind, nur für bestimmte Risikokategorien und für ein bestimmtes Risikoniveau zuzulassen oder eine Risikogewichtung vorzusehen, die mit Hilfe dieser Systeme berechnet wird und strengere Schwellenwerte als die in Basel III vorgesehenen vorsieht.

B. Praktische Empfehlungen für Bankinstitute

Ohne hier auf die Details zur Einrichtung eines umfassenden Managementsystems für KI-Tools nach Standards wie NIST AI RMF oder ISO 42001 einzugehen, stellen die folgenden Punkte Maßnahmen dar, mit denen der im Bericht vorgeschlagene Ansatz umgesetzt werden kann :

- Alle KI-Anwendungsfälle kartieren und nach ihrer Kritikalität und Erklärbarkeit klassifizieren.
- Verabschiedung einer internen Richtlinie, in der für die Risikokategorien geeignete Schwellenwerte für die Erklärbarkeit festgelegt werden.
- Einführung einer unabhängigen Überprüfung, einschließlich SHAP/LIME-Tests und regelmäßiger Validierung während des gesamten Lebenszyklus dieser Systeme.
- Schulung der Compliance-, Risiko- und IT-Teams zu den Anforderungen der FINMA und den Grenzen der möglicherweise verwendeten Techniken.
- Vorsehen von Massnahmen zur Minimierung der Risiken, die sich aus einer unzureichenden Erklärbarkeit ergeben.
- Durchführung eines simulierten internen FINMA-Konformitätsaudits zur Erklärbarkeit und KI-Governance.
- Dokumentation des Ansatzes der Bank in Bezug auf die Erklärbarkeit anhand konkreter Anwendungsfälle, die die Vorteile komplexer Systeme in Bezug auf Risiken und Kundenbelange aufzeigen, mit einer Begründung für die für bestimmte Systeme gewährten Ausnahmeregelungen und die Maßnahmen zur Minimierung der Risiken auf ein akzeptables Maß.

C. Schlussfolgerung

Der vom FSI vorgeschlagene Ansatz bietet einen möglichen Interpretationsrahmen, der es Bankinstituten ermöglicht, die mit dem Einsatz von KI-Systemen verbundenen Risiken unter Berücksichtigung ihres Erklärbarkeitsgrades zu bestimmen, zu bewerten und „verantwortungsvolle" Entscheidungen in dieser Angelegenheit zu treffen. Es geht also nicht mehr darum, zwischen Leistung und Erklärbarkeit zu wählen, sondern einen Governance-Rahmen zu schaffen, der es ermöglicht, komplexe Modelle mit einem angemessenen Kontrollniveau zu nutzen. Diese Entwicklung erfordert erhebliche Investitionen in Kompetenzen und Prozesse, könnte aber die Einführung von KI in kritischen Bankaktivitäten erheblich beschleunigen. Institute, die diese Entwicklungen vorwegnehmen, werden sich einen entscheidenden Wettbewerbsvorteil bei der digitalen Transformation des Sektors verschaffen.
