Imagem gerada por IA.

Confiança no código:
Governança e segurança da IA no setor financeiro brasileiro.

Artigo escrito por Maria Eva Mit Lazarin.

Minibio: Empreendedora e pesquisadora. Está a frente de empresas de tecnologia e investimento, preside o Conselho Consultivo da ABRIA - Associação Brasileira de Inteligência Artificial e conselheira do AI Safety Brasil.

Junho de 2026

A cibersegurança deixou de ser compreendida apenas como uma dimensão operacional da tecnologia bancária e passou a ocupar posição central na agenda de estabilidade financeira. Em sistemas financeiros altamente digitalizados, incidentes cibernéticos podem comprometer não apenas a disponibilidade de serviços, mas também a confiança dos agentes, a continuidade operacional das instituições e, em cenários extremos, a estabilidade macrofinanceira [1][2][3]. Esse deslocamento é particularmente relevante em um ambiente no qual instituições financeiras, fintechs, provedores de infraestrutura, plataformas de pagamento, serviços em nuvem, APIs e modelos de inteligência artificial passam a compor uma arquitetura tecnológica interdependente.

A inteligência artificial amplia esse debate porque transforma dados, modelos e infraestruturas digitais em componentes centrais da decisão financeira. No setor financeiro, sistemas algorítmicos já são utilizados em análise de crédito, concessão de limites, prevenção à fraude, precificação de risco, prevenção à lavagem de dinheiro, atendimento automatizado, gestão de portfólio e monitoramento de transações. A eficiência dessas aplicações é significativa, mas ela vem acompanhada de novos desafios: opacidade decisória, vieses algorítmicos, dependência de terceiros tecnológicos, riscos cibernéticos, falhas correlacionadas e dificuldades de responsabilização.

Antes de aprofundar a questão da governança e da segurança da IA no setor financeiro brasileiro, é necessário delimitar três pilares centrais dessa discussão: a explicabilidade algorítmica, a auditoria e validação de modelos e a segurança cibernética da infraestrutura de IA.

O primeiro pilar é a explicabilidade algorítmica, ou Explainable Artificial Intelligence — XAI —, compreendida como o conjunto de métodos, práticas e artefatos técnicos destinados a tornar decisões automatizadas mais compreensíveis, justificáveis, auditáveis e contestáveis [4][5]. No setor financeiro, a XAI deixa de ser apenas uma aspiração ética de transparência e passa a funcionar como requisito indireto de governança, conformidade, gestão de risco de modelo e supervisão prudencial [6][7]. Isso ocorre porque decisões financeiras automatizadas produzem efeitos concretos sobre indivíduos, empresas e mercados: um crédito pode ser negado, um limite pode ser reduzido, uma transação pode ser bloqueada, uma operação pode ser classificada como suspeita ou uma carteira pode ser rebalanceada por critérios que nem sempre são imediatamente compreensíveis.

Esse movimento é reforçado por marcos internacionais, como o AI Act da União Europeia, pelas diretrizes do High-Level Expert Group on Artificial Intelligence da Comissão Europeia e por documentos recentes de organismos como o BIS/FSI, o FSB, a OCDE e o FMI [1][3][6][8][9]. Esses referenciais não devem ser compreendidos apenas como normas externas, mas como espelhos regulatórios relevantes para bancos centrais e autoridades supervisoras em diferentes jurisdições. No Brasil, esse debate dialoga diretamente com a atuação do Banco Central, com a Resolução CMN nº 4.893/2021 sobre segurança cibernética e contratação de serviços de processamento, armazenamento e computação em nuvem, com a Resolução BCB nº 85/2021 aplicável às instituições de pagamento, com a LGPD e com a expansão de infraestruturas críticas como Pix, Open Finance e ecossistemas de fintechs [10][11][12].

O segundo pilar é a mitigação de vieses, a validação de modelos e a auditoria algorítmica. Instituições financeiras utilizam modelos cada vez mais complexos para processar grandes volumes de dados e produzir decisões em tempo real. Técnicas como redes neurais, modelos de boosting, Random Forest, XGBoost e sistemas baseados em aprendizado profundo podem elevar a acurácia preditiva, mas também dificultam a compreensão dos critérios efetivamente utilizados na decisão [4][5][13]. Esse é o problema clássico da "caixa-preta": modelos estatisticamente poderosos podem produzir decisões relevantes sem oferecer, de forma imediata, uma justificativa inteligível para reguladores, auditores, instituições e cidadãos afetados.

Nesse cenário, métodos como LIME e SHAP tornam-se instrumentos relevantes para traduzir a lógica de modelos complexos em critérios mais compreensíveis[14][15].

A auditoria algorítmica, entretanto, não se limita à geração de explicações. Ela exige documentação, rastreabilidade, testes independentes, monitoramento contínuo, análise de vieses, avaliação de robustez, revisão periódica e governança ao longo de todo o ciclo de vida do modelo. Frameworks como o RESHAPE, por exemplo, demonstram como técnicas de explicabilidade podem ser aplicadas à auditoria financeira, especialmente na identificação de anomalias contábeis e na explicação de padrões detectados por modelos complexos [16]. Ainda assim, esse tipo de ferramenta deve ser tratado como exemplo técnico específico, e não como solução geral para todos os desafios regulatórios da IA financeira.

O terceiro pilar é a segurança cibernética da infraestrutura de IA. A confiança em sistemas algorítmicos não depende apenas da qualidade estatística do modelo, mas também da integridade dos dados, da segurança dos ambientes de treinamento e inferência, da proteção contra ataques adversariais, da governança de terceiros, da rastreabilidade das decisões e da resiliência dos serviços críticos [1][2][17][18]. Se a integridade de um modelo ou de sua camada explicativa for comprometida, agentes maliciosos podem manipular entradas, explorar vulnerabilidades, induzir classificações equivocadas ou produzir explicações aparentemente legítimas para decisões incorretas. Em aplicações financeiras, esse risco pode afetar crédito, seguros, pagamentos, antifraude, gestão de liquidez, gestão de empréstimos e resposta a eventos de mercado.

Esse ponto é particularmente relevante porque as próprias ferramentas de explicação também podem se tornar vulneráveis. Explicações algorítmicas mal calibradas, incompletas, manipuláveis ou excessivamente simplificadas podem gerar uma falsa sensação de segurança [4][6]. Além disso, ataques de adversarial machine learning, data poisoning, model extraction, prompt injection, vazamento de dados e exploração de APIs podem comprometer não apenas o modelo principal, mas também os mecanismos usados para justificá-lo [17][18]. Portanto, a explicabilidade precisa ser acompanhada por controles de segurança, validação independente e governança da infraestrutura técnica que sustenta a IA.

O risco sistêmico emerge quando falhas individuais deixam de ser isoladas e passam a ser correlacionadas. Se múltiplas instituições financeiras utilizam modelos semelhantes, bases de dados concentradas, provedores comuns de nuvem, APIs compartilhadas, LLMs de terceiros ou arquiteturas de IA opacas, uma mesma vulnerabilidade técnica pode produzir efeitos simultâneos em diferentes agentes do sistema financeiro [1][2][19]. O problema, portanto, não está apenas na “caixa-preta” de um modelo específico, mas na possibilidade de que a opacidade, a dependência tecnológica, a concentração de fornecedores e a automação de decisões criem canais de propagação de choques.

Os sistemas de IA podem tanto estabilizar quanto amplificar crises financeiras, dependendo da forma como os modelos reagem a choques, da presença de comportamentos endógenos, da existência de

complementaridades estratégicas e dos objetivos atribuídos aos sistemas automatizados [19]. Essa perspectiva permite compreender que o risco da IA no setor financeiro não é apenas microprudencial, associado à falha de uma instituição específica, mas também macroprudencial, associado à possibilidade de respostas homogêneas, falhas simultâneas e propagação sistêmica.

No Brasil, esse debate ganha relevância particular em razão da sofisticação tecnológica do ecossistema financeiro, da expansão das fintechs, da centralidade dos pagamentos instantâneos, da digitalização do crédito, do avanço do Open Finance e da crescente dependência de infraestruturas de dados e computação em nuvem. A IA já participa de decisões relevantes no mercado financeiro brasileiro, desde a análise de crédito em grandes bancos até a concessão de limites em fintechs e os sistemas antifraude que monitoram transações em tempo real. A eficiência dessas aplicações é significativa, mas a sua legitimidade depende da capacidade de explicar, auditar, proteger e contestar decisões automatizadas [3][5][6][7].

Assim, “confiança no código” no setor financeiro brasileiro deve ser compreendida como uma arquitetura regulatória e técnica composta por três camadas interdependentes: governança algorítmica, explicabilidade auditável e resiliência cibernética da infraestrutura de IA. A IA financeira não cria apenas o risco individual de uma decisão incorreta; ela pode criar risco sistêmico quando modelos semelhantes, dados concentrados, provedores críticos, APIs, LLMs e infraestruturas compartilhadas produzem falhas correlacionadas [1][2][19]. A confiança, portanto, não decorre apenas da acurácia dos modelos, mas da capacidade institucional de documentar, auditar, proteger, monitorar e responsabilizar sistemas algorítmicos ao longo de todo o seu ciclo de vida.

Nesse sentido, o desafio regulatório contemporâneo não é impedir a inovação financeira baseada em IA, mas estabelecer condições para que ela seja segura, explicável, auditável e resiliente. O setor financeiro brasileiro dispõe de uma base tecnológica avançada e de uma autoridade regulatória reconhecida por sua capacidade de inovação institucional. O próximo passo é integrar a governança de IA à agenda prudencial, cibernética e sistêmica, reconhecendo que, em mercados cada vez mais automatizados, a confiança não está apenas nas instituições: ela também está no código que decide, recomenda, bloqueia, aprova, precifica e executa.

REFERÊNCIAS

[1] Financial Stability Board — FSB. The Financial Stability Implications of Artificial Intelligence. 2024.

[2] Aldasoro, I.; Gambacorta, L.; Giudici, P.; Leach, T. The drivers of cyber risk. Journal of Financial Stability, 2022. DOI: 10.1016/j.jfs.2022.100989.

[3] OECD. Regulatory approaches to Artificial Intelligence in finance. 2024.

[4] Arrieta, A. B. et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020.

[5] Černevičienė, J.; Kabašinskas, A. Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 2024. DOI: 10.1007/s10462-024-10854-8.

[6] Bank for International Settlements / Financial Stability Institute — BIS/FSI. Managing explanations: how regulators can address AI explainability. 2025.

[7] Giudici, P.; Raffinetti, E. SAFE Artificial Intelligence in finance. Finance Research Letters, 2023. DOI: 10.1016/j.frl.2023.104088.

[8] European Union. Regulation (EU) 2024/1689 — Artificial Intelligence Act. 2024.

[9] High-Level Expert Group on Artificial Intelligence — HLEG. Ethics Guidelines for Trustworthy AI. European Commission, 2019.

[10] Conselho Monetário Nacional — CMN. Resolução CMN nº 4.893/2021. Dispõe sobre política de segurança cibernética e requisitos para contratação de serviços de processamento, armazenamento de dados e computação em nuvem.

[11] Banco Central do Brasil — BCB. Resolução BCB nº 85/2021. Dispõe sobre política de segurança cibernética e requisitos para contratação de serviços por instituições de pagamento.

[12] Brasil. Lei Geral de Proteção de Dados Pessoais — LGPD, Lei nº 13.709/2018.

[13] Weber, P. et al. Applications of Explainable Artificial Intelligence in Finance: a systematic review of Finance, Information Systems and Computer Science literature. Management Review Quarterly, 2024.

[14] Ribeiro, M. T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. Proceedings of KDD, 2016.

[15] Lundberg, S. M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems — NeurIPS, 2017.

[16] Müller, R.; Schreyer, M.; Sattarov, T.; Borth, D. RESHAPE: Explaining Accounting Anomalies in Financial Statement Audits by enhancing SHapley Additive exPlanations. 2022.

[17] National Institute of Standards and Technology — NIST. Artificial Intelligence Risk Management Framework — AI RMF 1.0. 2023.

[18] National Institute of Standards and Technology — NIST. Cybersecurity Framework 2.0. 2024.

[19] Danielsson, J.; Uthemann, A. Artificial Intelligence and Financial Crises / On the use of artificial intelligence in financial regulations and the impact on financial stability. Systemic Risk Centre / LSE, 2023–2024.