3 Ways To Limit Risks Of Black-Box AI In Financial Services

By **Jeffrey Naimon and Caroline Stapleton** (August 21, 2024)

Jeffrey Naimon

Caroline Stapleton

Acting Comptroller of the Currency Michael Hsu delivered remarks earlier this summer on the risks of artificial intelligence, including "the black box nature of AI and what that means for accountability and risk governance."[1]

Hsu's June 6 remarks highlighted the use of AI in credit decisioning and noted the tension between AI's potential for expanding access to credit and the opacity of its underlying algorithms:

> For those who would have been denied by the AI algorithm, there is a question of fairness. Why was I denied? Data sets can be biased, algorithms can hallucinate, and reinforcement learning from human feedback can yield mistakes. How can one trust that the decisions reached by an AI algorithm are fair?[2]

The acting comptroller is not the first official to express concern about the black-box problem that financial institutions confront when adopting AI.

The Consumer Financial Protection Bureau, the U.S. Department of the Treasury and other regulators have weighed in on the importance of ensuring explainability and transparency when companies leverage AI models to offer credit and other financial products to consumers.[3] The White House has also emphasized the importance of providing understandable explanations "as to how and why a decision was made or an action was taken by an automated system."[4] And regulators at the state and federal level are doubling down on disclosure requirements when AI is used in credit and other significant decisions affecting consumers.[5]

As the concept of AI explainability increasingly takes center stage, what should financial institutions be doing to ensure they provide the transparency that their regulators and customers expect — and in some cases, require? Below are considerations for companies seeking to mitigate the regulatory and compliance risk presented by black-box AI tools in financial services.

1. Has your institution created an AI inventory?

Interagency guidance on model risk management adopted by the Office of the Comptroller of the Currency, Federal Reserve and Federal Deposit Insurance Corp. states that banks "should maintain a comprehensive set of information for models implemented for use, under development for implementation, or recently retired."[6]

According to the guidance, an inventory should include, among other things, the model's purpose, the products for which it is used, the types and sources of inputs, and the individuals responsible for overseeing the model's implementation and use.[7]

The Comptroller's Handbook on Model Risk Management further notes that sound AI risk management typically requires an inventory of AI uses.[8] Creating and updating an inventory of existing, expected or proposed AI uses can help institutions evaluate their black-box risk by identifying models that are inherently opaque, and for which the guidance indicates that regulators expect creditors to undertake additional oversight and scrutiny as a result.

Developing an AI inventory may also reveal where AI and non-AI models are being used for the same or similar products, services or functions — which can help institutions determine whether a black-box model is actually necessary to perform a given task. For example, an inventory may reveal that a black-box AI model is used in an initial step of credit underwriting, with a non-AI model or manual process used as an overlay.

This structure could prompt the institution to assess the additional value that the AI model is providing in the underwriting process and to determine whether this added value outweighs the black-box explainability risk inherent in many AI models.

2. Does — or will — your institution use AI to make decisions that trigger explanatory notice requirements?

As noted above, creating an AI inventory can clarify the functions where AI is used — or may be used in the future — to offer products, communicate with consumers or terminate services.

Institutions may look to the inventory to identify which of their AI uses may be subject to a legal explainability requirement, such as adverse action notice requirements under the Equal Credit Opportunity Act or Fair Credit Reporting Act, or a state AI disclosure mandate, such as that signed into Colorado law in May. This is important because using AI in connection with activities that trigger notice requirements may necessitate changes in the way that disclosures are generated and tested for regulatory compliance.

The CFPB has been particularly vocal about the impact of AI on compliance with adverse action notice requirements. In circulars published in 2022 and 2023, the CFPB warned lenders that using complex algorithms in credit decisioning does not change their obligation to provide a compliant adverse action notice that discloses "the factors actually considered or scored by the creditor," even if these factors "may be surprising to consumers."[9]

According to a May 2022 CFPB circular, if an AI tool used in credit decisions is so complex or opaque that a lender "cannot provide the specific and accurate reasons for adverse actions," the CFPB likely would find its use violates the Equal Credit Opportunity Act.[10] This broadly discouraging perspective is especially interesting because many creditors using AI models are doing so with the intention of identifying good-credit consumers whose traditional credit scores might place them outside most creditors' standard criteria.

In light of this regulatory guidance, financial institutions should confirm they are able to consistently and accurately identify the principal explanatory factors for credit decisions made using AI. Solutions may include, for example, developing a separate model to identify reason codes for credit denials using Shapley values or other accepted methodologies.[11]

In addition, entities should consider whether the language used to communicate these factors in an adverse action notice is meaningful and digestible by consumers.[12] The CFPB and other regulators are more likely to scrutinize adverse action reasons that are vague, confusing or not intuitive to consumers.

3. How will your institution assess and monitor AI on an ongoing basis?

Regulators expect financial institutions to perform periodic risk assessments of AI models. As the Comptroller's Handbook expressly suggests, this includes an assessment methodology that considers AI model explainability.[13] For more complex models, or those that learn and modify their parameters over time, the black-box risk may be heightened and warrant additional preimplementation testing and postimplementation monitoring and validation.

In addition, ongoing oversight is an essential component of model risk management, including AI models. While it is vital to perform validation of the AI model itself, it may be necessary for highly complex or opaque models to validate not only that the model's outputs are satisfactory and performing as expected, but that the reasons for the outputs continue to be discernible, understandable and, where necessary, disclosed to the appropriate audiences.

For example, are there variables that serve as model inputs, but are rarely or never identified as explanatory factors for model outputs? If so, it may be prudent to determine whether these variables actually lack explanatory power, or whether the explanation methodology is failing to detect their significance.

While these considerations may serve as a starting point for mitigating black-box risk, the best framework for assessing and validating the explainability of AI models will depend on each institution's specific use case. Lenders and other financial services providers should consult with model developers — whether internal or third parties — and reviewers to ensure that explainability is a component of the entity's overall AI risk management strategy.

Jeffrey Naimon and Caroline Stapleton are partners at Orrick Herrington & Sutcliffe LLP.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

[1] Acting Comptroller Hsu, Remarks on AI Tools, Weapons, and Accountability: A Financial Stability Perspective, at *10 (June 2024); https://www.occ.gov/news-issuances/speeches/2024/pub-speech-2024-61.pdf.

[2] Id.

[3] See e.g., CFPB et al., Joint Statement on Enforcement of Civil Rights, Fair Competition, Consumer Protection; https://www.justice.gov/crt/media/1346821/dl?inline, and Equal Opportunity Laws in Automated Systems (Apr. 2024); U.S. Treasury Department, Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector: https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf, at *36-37 (Mar. 2024).

[4] White House, Blueprint for an AI Bill of Rights, at *43-44 (Oct. 2022).

[5] See e.g., CFPB, Consumer Financial Protection Circular 2023-03; https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-

action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/, Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B (Sept. 2023); CFPB, Consumer Financial Protection Circular 2022-03; https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/, Adverse action notification requirements in connection with credit decisions based on complex algorithms (May 2022); Colo. Rev. Stat. § 6-1-1701 et seq.; https://leg.colorado.gov/sites/default/files/2024a_205_signed.pdf, (recently enacted algorithmic accountability law that requires, in certain circumstances, a notice explaining the reason(s) for a consequential decision made using AI). Interestingly, regulators have spilled relatively less ink on whether AI models actually function for their intended purposes, typically in predicting the likelihood of default to help drive credit and pricing decisions.

[6] Federal Reserve, SR Letter 11-7, Supervisory Guidance on Model Risk Management, at *20-21 (2011); https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf.

[7] Id.

[8] Comptroller's Handbook, Model Risk Management, at *13 (Aug. 2021); https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.

[9] CFPB, Consumer Financial Protection Circular 2023-03; https://www.consumerfinance.gov/compliance/circulars/circular-2023-03-adverse-action-notification-requirements-and-the-proper-use-of-the-cfpbs-sample-forms-provided-in-regulation-b/, Adverse action notification requirements and the proper use of the CFPB's sample forms provided in Regulation B (Sept. 2023); see also CFPB, Consumer Financial Protection Circular 2022-03; https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/, Adverse action notification requirements in connection with credit decisions based on complex algorithms (May 2022).

[10] CFPB, Consumer Financial Protection Circular 2022-03; https://www.consumerfinance.gov/compliance/circulars/circular-2022-03-adverse-action-notification-requirements-in-connection-with-credit-decisions-based-on-complex-algorithms/, Adverse action notification requirements in connection with credit decisions based on complex algorithms (May 2022).

[11] See, e.g., Pace, Using Explainable AI to Produce ECOA Adverse Action Reasons: What Are The Risks? At https://www.paceanalyticsllc.com/post/ecoa-adverse-actions-and-explainable-ai.

[12] See National Institute of Standards and Technology, Four Principles of Explainable Artificial Intelligence; https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf, at *ii (Sept. 2021) (articulating four principles of AI explainability, including meaningfulness of the explanation).

[13] Comptroller's Handbook on Model Risk Management, at *91 (Aug. 2021); https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/pub-ch-model-risk.pdf.