truera

Financial Institutions' Use of Artificial Intelligence, Including Machine Learning

REQUEST FOR INFORMATION RESPONSE

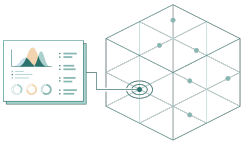
for the

US Department of Treasury
Office of the Comptroller of the Currency
Federal Deposit Insurance Corporation
Consumer Financial Protection Bureau
and the **National Credit Union Administration**

Docket No. OCC-2020-0049

About Us

TruEra provides AI Quality solutions that analyze machine learning, drive model quality improvements, and build trust. Powered by enterprise-class Artificial Intelligence (AI) Explainability technology based on six years of research at Carnegie Mellon University, TruEra's suite of solutions provides much-needed model transparency and analytics that drive high model quality and overall acceptance, address unfair bias, and ensure governance and compliance.



Authors

Anupam Datta

Co-Founder, President, Chief Scientist

Anupam is passionate about enabling responsible adoption of artificial intelligence. As a Professor of Electrical and Computer Engineering and Computer Science at Carnegie Mellon University for over a decade, he has led groundbreaking research in the areas of AI explainability and governance as well as privacy and data protection. Anupam obtained PhD and MS degrees from Stanford University and a BTech from the Indian Institute of Technology, Kharagpur, all in Computer Science.

Shameek Kundu

Chief Strategy Officer, Head of Financial Services

Shameek has spent most of his career in driving responsible adoption of data analytics/ AI in the financial services industry. He is a member of the Bank of England's AI Public-Private Forum and the OECD Global Partnership on AI, and was part of Monetary Authority of Singapore's Steering Committee on Fairness, Ethics, Accountability and Transparency in AI. Most recently, Shameek was Group Chief Data Officer at Standard Chartered Bank, where he helped the bank explore and adopt AI in multiple areas, and shaped the bank's internal approach to responsible AI.

Divya Gopinath

Research Engineer

Divya is a research engineer with expertise in machine learning and full-stack development. She holds Bachelor's and Master's degrees in Computer Science and Engineering from the Massachusetts Institute of Technology, where her research focused on building interpretable machine learning algorithms currently deployed to help doctors manage emergency room patients.

Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, Including Machine Learning

Key takeaways

(1) **Explainability should be the backbone of assuring high-quality and trustworthy AI/ML models.** It should be embedded throughout the model lifecycle from development to validation to continuous monitoring in production.

(2) **We expect both inherently/structurally interpretable models *and* algorithmically interpretable ones (with post-hoc explanation methods) to co-exist.** We recommend viewing these approaches to explainability as complementary and reinforcing each other rather than a mutually exclusive choice. Constraining the industry to use a small set of models that are viewed as inherently explainable (e.g., linear models and Bayesian rulesets) could have a detrimental impact on innovation. Instead, it would be better to identify a set of questions about models that should be answered as part of explainability and model risk management activities and map them to appropriate technical tools and processes to provide responsible governance.

(3) **We believe that the current 'state of the art' in post-hoc explanation methods, if applied correctly, can provide explanations that are accurate enough to meet customer and regulatory expectations adequately.** However, this requires significant focus by Financial Institutions (FIs) and their partners on appropriate design and implementation of their explanation methodology.

(4) **Explanation outputs can dramatically and meaningfully differ based on the comparison group and the output type that is being explained.** FIs - particularly those with extensive adoption of AI/ML - may want to consider introducing internal standards around the accuracy and consistency of explanation methods.

(5) **In order to ensure compliance with ECOA/Regulation B, regulators should consider providing greater clarity on appropriate measures of fairness for specific use cases (beyond credit).** FIs should consider employing root cause analysis to understand and justify any sources of bias before declaring a model "fair," or otherwise.

1. How do financial institutions identify and manage risks relating to AI explainability? What barriers or challenges for explainability exist for developing, adopting, and managing AI?

When it comes to AI explainability, FIs have a choice of two approaches: use only those models that are structurally/inherently explainable, or use algorithmically explainable models that are supported by post-hoc explainable models.

Each has its own challenges.

- Constraining the use of AI/ML to inherently explainable models alone can prevent the industry from capturing the full value from AI/ML, given that other models can provide significantly better predictive accuracy in some areas ***(Takeaway 2).*** In the extreme, this could put U.S. financial firms and AI technology companies at a competitive disadvantage versus companies in other regulatory jurisdictions.
- On the other hand, not all implementations of post-hoc explainability methods can provide reliable expectations. The current state of the art in post-hoc explanation methods, if applied correctly, can provide explanations that are accurate enough to meet customer and regulatory expectations adequately. However, FIs need to pay significant attention to their design and implementation to ensure reliability ***(Takeaway 3).***

In practice, we believe that the two approaches will co-exist and complement each other. The sheer range of AI/ ML use cases in the industry, and the differing levels of maturity of the two approaches in many of these areas, would suggest the need for such coexistence. For example, alternatives to inherently explainable models perform better in image recognition[1] and natural language processing[2]. Even for more traditional financial industry models, gradient-boosted tree ensembles and neural networks outperform traditional tree and linear models in their predictive accuracy.

Furthermore, the boundary of which models are inherently interpretable is shifting. Models that were at one point not viewed as inherently interpretable (e.g., ReLU networks with a few layers) are now being considered so by some researchers. We expect this trend to continue as the community creates better tools for understanding complex models.

---

[1] "[1903.09190] Comparison of State-of-the-Art Deep Learning ... - arXiv." 11 Jun. 2020, https://arxiv.org/abs/1903.09190. Accessed 19 May. 2021.
[2] "BERT: Pre-training of Deep Bidirectional Transformers for Language ...." https://arxiv.org/abs/1810.04805. Accessed 19 May. 2021.

The real challenges in using explainability to help scale up AI adoption lie elsewhere - in the narrow way in which explainability has been used in many parts of the industry so far, and the lack of alignment on the objectives of explainability.

- **How explainability is used within the AI/ML model lifecycle:** To date, explainability has largely been relegated to a narrow part of the model development life cycle, where model builders or validators might examine a model's interpretability prior to recommending it for deployment. However, explainability can actually enable trust in a model through its life cycle, from iterative development to analyzing it in terms of fairness, stability, etc. and even monitoring a model's stability once it is deployed ***(Takeaway 1)***.
- **Explainability objectives:** Should explanations enable FIs to justify - internally and to regulators - how they have arrived at decisions? Or are they intended to educate and provide redress options to customers impacted by decisions? And should the bar on these expectations be calibrated based on materiality or other factors? We believe that clearer alignment on what queries about model decisions need to be explained and what explanation techniques are expected for use cases of varying materiality will be beneficial **.**

The science around explanation of AI/ML models continues to evolve. Given the wide range of explanation methods available, FIs - particularly those with extensive adoption of AI/ML - may want to consider introducing internal standards around the accuracy and consistency of explanation methods. ***(Takeaway 4)***

2. How do financial institutions use post-hoc methods to assist in evaluating conceptual soundness? How common are these methods? Are there limitations of these methods (whether to explain an AI approach's overall operation or to explain a specific prediction or categorization)? If so, please provide details on such limitations.

A good explanation method must:

- **Be accurate enough** to meet customer and regulatory expectations, though the accuracy standards set by an FI might vary based on materiality
- **Support both local and global explanations**, providing justification for a specific individual's decision (needed for adverse action notices and fine-grained debugging of a model) as well as identifying the aggregate drivers of model predictions (needed for conceptual soundness checks for both data scientists and risk management teams).
- **Be able to answer a rich set of queries**. Explaining a classification decision, for example ("Why was Jane denied a loan?") is fundamentally different from

explaining a risk score ("Why was Jane assigned a risk score of 0.6?")[3]. This can also depend on the comparison group-- "Why was Jane denied a loan compared to all approved applicants?" might yield different results than "Why was Jane denied a loan compared to all approved applicants in her income bracket?". Explaining the reasons driving a decision is different from providing guidance on the actions that a customer can take to improve their decision outcome.

- **Capture causal influence** between inputs and the final model outcome.

A natural choice of explanation techniques that satisfy these criteria are those that compute the Shapley value, which is a notion that borrows from cooperative game theory to capture the average marginal contribution of a feature to a model outcome. There are a variety of methods that approximate Shapley values including Shapley Additive exPlanations (SHAP)[4] and Quantitative Input Influence (QII)[5]. There are analogues of these methods for gradient-based models like neural networks, such as Integrated Gradients[6]. Counterfactual explanations provide a useful complementary approach, when it comes to providing *actionable* guidance to customers on improving their decision outcomes.

Many FIs that have AI/ML models in production have begun using some form of post-hoc explainability. When it comes to the limitations of such methods, it is useful to think of two different aspects: limitations that are inherent to such methods, and those that are a function of poor implementation.

- Post-hoc explanation methods are, by definition, estimates. However, post-hoc methods such as those using Shapley values can already be designed and implemented to provide acceptable levels of accuracy in explaining the key drivers of decisions (e.g., accuracy guarantees of the form "X% of exact values Y% of the time"). When combined with more detailed feature influence plots, they should provide FIs the necessary level of confidence in their AI/ ML models.
- The more serious challenge relates to the quality and consistency of implementation of such explanation methods - which can result in significantly different levels of explanation accuracy. There is also a need to look at the way in which explanations are communicated: the expectations of internal stakeholders may well differ from those of customers aggrieved by an unfavourable decision.

---

[3] "Machine learning models require the right explanation framework ...." https://truera.com/machine-learning-models-require-the-right-explanation-framework-and-its-easy-to-get-wrong/. Accessed 19 May. 2021.
[4] "A Unified Approach to Interpreting Model Predictions." 22 May. 2017, https://arxiv.org/abs/1705.07874. Accessed 19 May. 2021.
[5] "Algorithmic Transparency via Quantitative Input Influence:." https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf. Accessed 19 May. 2021.
[6] "Axiomatic Attribution for Deep Networks." 4 Mar. 2017, https://arxiv.org/abs/1703.01365. Accessed 19 May. 2021.

There have also been concerns raised about algorithmically generated interpretations of model outputs being susceptible to adversarial attacks and thereby producing strange explanations[7, 8]. However, there is growing consensus in the literature that this is less a byproduct of the explanation method and more illustrative of the model's weakness/ instability[9] or even an unrealistic adversarial attack model. There is also evidence that a model that is trained robustly yields explanations that are much less susceptible to these attacks[10]. Robustness is a generally desirable property for models regardless of the explanation method used. A good explanation technique will not be able to overcome the limitations of a model that is poorly designed or trained.

3. For which uses of AI is lack of explainability more of a challenge? How do institutions account for and manage the varied challenges and risks posed by different uses?
While explainability is arguably important for all AI use cases, it becomes particularly important wherever the algorithm's recommendation/prediction could impact:

- Customer access to financial services (e.g., Know Your Customer/client risk rating, insurance cover, loan), particularly for 'natural person' customers
- An FI's ability to meet its fiduciary obligations to customers (e.g., investment advice, best execution of trades)
- An FI's ability to meet regulatory obligations, either prudential (e.g., capital, liquidity or market risk calculations) or conduct-wise (e.g., financial crime compliance, market competition and stability)

FIs are at different levels of maturity when it comes to addressing these challenges. Emerging best practice seems to involve the following elements:

- A materiality scale that aligns with the FI's broader risk management framework, and can be used to determine different minimum levels of explainability for different use cases.
- Embedding of explainability not just at the end of the development process or during validation, but across the end to end model lifecycle. This allows transparency into the model at all stages of development and ensures that the final product meets risk management standards; as an example, we later talk about how explainability can inform fairness assessments.

---

[7] "Towards Robust Explanations for Deep Neural Networks." 18 Dec. 2020, https://arxiv.org/abs/2012.10425. Accessed 19 May. 2021.
[8] "Explanations can be manipulated and geometry is to blame." 19 Jun. 2019, https://arxiv.org/abs/1906.07983. Accessed 19 May. 2021.
[9] "Fooling Neural Network Interpretations via Adversarial Model ...." 6 Feb. 2019, https://arxiv.org/abs/1902.02041. Accessed 19 May. 2021.
[10] "Smoothed Geometry for Robust Attribution." 11 Jun. 2020, https://arxiv.org/abs/2006.06643. Accessed 19 May. 2021.

- Recognition of the differences between internal and external (customer-facing) explainability requirements. There is relatively limited research on the latter.
- Flexibility in the choice of model type and explainability techniques, while specifying certain minimum standards.
- Effective technology support to enable the right level of explainability.

4. How do financial institutions using AI manage risks related to data quality and data processing? How, if at all, have control processes or automated data quality routines changed to address the data quality needs of AI? How does risk management for alternative data compare to that of traditional data? Are there any barriers or challenges that data quality and data processing pose for developing, adopting, and managing AI? If so, please provide details on those barriers or challenges.

Many of the challenges around data quality and data processing for AI are not unique to AI. Indeed, the availability of, and access to, data of the right quality has been an area of focus for nearly two decades now.

However, because of its data hungry nature, AI does create incremental challenges in the following aspects of data quality management

- **'Representativeness' of training data.** While some argue that the existing 'completeness' principle of most data governance frameworks covers this, we believe that it is inadequate. Traditional considerations of data completeness inside an FS has focused on ensuring that all the data available to the FI and relevant to a use case is brought to bear. However, when it comes to training data for AI, that is necessary but not sufficient - the FI might simply lack the data needed to ensure that the AI model is trained with a data set representative of the population on which the model must operate in the future.
- **Moving away from absolute measures of data quality to 'DQ scores':** In conventional thinking around data quality, all exceptions matter and thresholds for missing or potentially incorrect data can be very low in use cases like Financial Crime Compliance. However, if those same exacting thresholds were applied to a large data set used to train an AI model and containing hundreds of input features, very few AI models would get built! Working out the appropriate tolerance levels for missing or potentially incorrect data when used for training AI models is arguably one of the biggest AI-related challenges for Chief Data Officers.
- **Explainability-linked DQ monitoring:** As a logical extension of the above, DQ monitoring also needs to be more industrialized once an AI model goes into production. Not every feature will matter equally to the final outcome of a model. FIs will benefit from prioritizing their scrutiny of ongoing DQ issues on those features that are known to make a material difference to the final outcome.

Beyond DQ, AI also introduces additional challenges on other aspects of data management and governance, including the retention of different model versions that are linked to potentially disparate training/test data sources, as well as privacy concerns such as ensuring that the AI model does not use protected variables, and that redress options or counterfactual explanations that are provided to customers do not breach others' privacy rights.

Finally, the use of third party or "alternate" data sources (e.g., telecom or social media for credit decisioning) creates an additional set of data management challenges:

- When a third party provides a calculated score rather than raw data, this can perpetuate the algorithmic black box model. The third party will often be unwilling to share their 'secret sauce' (whether using AI or otherwise) to the FI, but this leaves the FI exposed to not knowing how one of the inputs into *their* automated decision making is calculated.
- Managing consent in a reliable, scaleable and ideally automated manner when dealing with third party data (e.g., how do you ensure that when an individual withdraws consent for the use of their personal data from the first party, the FI using that data through a data broker is made aware and can automatically stop using it?)
- Use-case specific regulatory constraints on using data elements: certain jurisdictions ban the use of specific personal data elements for specific types of use cases (e.g., permitted for fraud detection but not for credit decisioning).
- Quality of data matching between internal and external data sets: without robust controls in this space, cases of 'mistaken identity' can occur. For example, an FI's record of their customer could be wrongly matched to another individual's data sourced from external sources, due to incorrect entity resolution.
- Segregation of internal/external data: without adequate thought, it may become impossible to segregate data sourced by an FI from their customers and transactions with that sourced from third parties. This means that should there be a reason to delete the external data in future (e.g., due to the discovery of inappropriate consent management by the third party), the FI may find it very difficult to execute. Even if deletion is possible, it might lead to the original model collapsing due to non availability of specific data elements.

Most FIs appear to be in early stages of incorporating these types of changes into their data management frameworks. This is not surprising, given the complexity of the issues at hand and the fact that many are in early stages of working through these implications.

5. Are there specific uses of AI for which alternative data are particularly effective?

One area where alternative data has certainly helped is the provision of credit to under-served customer segments. This is particularly the case in geographies that lack comprehensive credit bureaux, and/or penetration of traditional financial services is low.

6. How do financial institutions manage AI risks relating to overfitting? What barriers or challenges, if any, does overfitting pose for developing, adopting, and managing AI? How do financial institutions develop their AI so that it will adapt to new and potentially different populations (outside of the test and training data)?
Overfitting, and more broadly a lack of generalizability to new inputs or situations, is a critical barrier for the adoption and management of AI. Unstable models can be dangerous in production settings, but there are ways of preventing this. First, model transparency can once again be a backbone of ensuring a model's stability; by analyzing how drivers of a model's decision (both locally and globally) shift over time, we can determine whether the model has learned reasonable relationships between inputs and outputs and whether it is suited for new populations. Continuous monitoring of deployed AI/ML models is needed to track this over time, as we comment in the next question.

As stated in Takeaway 1, explainability is key to debugging overfitting, and these checks should be performed and monitored while the model is being developed, validated, and used in production.

8. How do financial institutions manage AI risks relating to dynamic updating? Describe any barriers or challenges that may impede the use of AI that involve dynamic updating. How do financial institutions gain an understanding of whether AI approaches producing different outputs over time based on the same inputs are operating as intended?

The lifecycle of a model does not end as soon as it is deployed-- continuous and ongoing monitoring of any production-ready model is critical. If unchecked, AI models may be unstable in new circumstances due to lack of generalizability or by operating on new data regimes. It is clear that human oversight of the model monitoring process is critical, and also that it is insufficient to only look at top-line accuracy metrics to determine whether a model is doing well in production. Analysis of *why* a model is making a particular decision and if this is shifting over time is also critical.

Put another way, as we assert in Takeaway 1, model transparency and explainability does not end after a model passes a risk assessment and makes it into production-- any thorough monitoring solution must examine not just underlying data drift but also how a model's explanations for decisions drift over time. Is the model operating in new territory but relying on a particular feature in a reasonable way? Is a model failing to properly

incorporate a feature value in its decision because it hasn't encountered extreme values before? These questions, among others, are important to investigate.

10. Please describe any particular challenges or impediments financial institutions face in using AI developed or provided by third parties and a description of how financial institutions manage the associated risks. Please provide detail on any challenges or impediments. How do those challenges or impediments vary by financial institution size and complexity?

AI developed or provided by third parties can typically be used by FIs in three ways:

1. As part of a broader system - e.g., financial crime investigation, fraud detection - that is implemented inside the FI
2. As a piece of insight - e.g., an alternative credit score, a 'sentiment marker' for individual securities - that is made available to the FI (without the underlying systems or models that helped come up with the insight)
3. As a service consumed by the FI - e.g., contract review conducted by a legal or professional services firm on behalf of the FI, but using an AI model for at least part of the task

The challenges differ across these categories. In the first category, the challenges can be largely addressed through traditional IT processes around vendor management and technology standards. There is some incremental awareness needed to ensure that those buying understand the system's use of AI and are comfortable with the level of transparency and AI quality assurance provided by the vendor.

In the second category, the challenge is arguably the most serious. It is often not obvious to the person buying such a score or insight that the vendor has created that insight using AI. As a result, most might not even realise that existing/ emerging guidelines around AI transparency and quality will apply - *to them and not just the vendor* - in these cases as well. Even if they understand the requirement, their ability to have the necessary oversight over the third party can be limited.

Similar risks also apply to the third category. However, because there is generally greater scrutiny in the procurement of such services anyway, both the awareness and oversight problem is arguably less significant in these instances.

11. What techniques are available to facilitate or evaluate the compliance of AI-based credit determination approaches with fair lending laws or mitigate risks of noncompliance? Please explain these techniques and their objectives, limitations of those techniques, and how those techniques relate to fair lending legal requirements.

There are two main camps of fairness: group fairness and individual fairness. Group fairness examines aggregate properties of segments of the population and seeks to equalize outcomes across these groups[11]. On the other hand, individual fairness enforces fairness at a more granular level, by asking that similar individuals should be treated similarly[12].

For group fairness, AI-based credit determination approaches should take a four-step approach to be compliant with fair lending laws: (1) determine the appropriate measure of fairness of the given problem; (2) evaluate their model with respect to this agreed upon metric; (3) perform a root cause analysis to determine the source(s) of any bias and whether or not this bias is justified; and (4) find a suitable mitigation strategy to correct this bias if unjustified.

Each step requires careful consideration. The chosen metric of fairness is reflective of what "fairness" means in a larger societal context as well as what information can be used to measure this fairness. As an example, if an ML model is being used to determine credit-worthiness, it only observes fair outcomes for applicants that the model determines deserve a loan-- if an applicant is given a loan, we then are able to observe whether or not they defaulted. However, if an individual is denied a loan, we cannot conclude that we have fair labels, because we do not know whether they would have truly defaulted in practice. In this way, our measure of fairness should strive to enact fair outcomes for the labels we can observe.

Even more importantly, step (3) is often skipped over, which can be dangerous. Understanding the source of a bias is critical to choosing a mitigation strategy that makes sense, and more fundamentally, determining if a bias is justified. Do we observe outcomes that in aggregate benefit men because the model is relying on income in a reasonable way and our population happens to have more men of a high income level than women? Or is the model truly building a proxy of gender? The answer to these questions has vast implications on whether bias is justified and needs to be removed to consider a model compliant with fair lending law. In this way, looking into the disparities between groups in terms of how a model uses a feature will guide mitigation and understanding of what went wrong.

---

[11] "NIPS 2017 Tutorial - Fairness in Machine Learning - Moritz Hardt." https://mrtz.org/nips17/. Accessed 19 May. 2021.
[12] "Fairness Through Awareness." 20 Apr. 2011, https://arxiv.org/abs/1104.3913. Accessed 19 May. 2021.

For individual fairness, on the other hand, model explainability is once again critical. Recall the Apple card debacle[13] where women claimed that their credit card applications were rejected while their husbands, who had identical financial profiles, were accepted. This is an example of individual fairness-- by answering questions like "Why was Jane denied a loan while her husband John was accepted?", we can drill into drivers of Jane's rejection compared to John's acceptances, and understand whether these decisions are truly justified. Individual fairness is often overlooked by FIs in favor of group fairness, but is equally critical to ensuring a model is fair.

In summary, as per Takeaway 5, FIs need clarity on appropriate measures of group and individual fairness in financial settings. FIs must also use root cause analysis to understand and justify sources of bias as a part of their standard fairness workflow.

12. What are the risks that AI can be biased and/or result in discrimination on prohibited bases? Are there effective ways to reduce risk of discrimination, whether during development, validation, revision, and/ or use? What are some of the barriers to or limitations of those methods?

AI/ML models can certainly be biased, either by encoding unjust biases that are present within training data (either due to historial prejudice or measurement error), or even by amplifying bias[14] from the data by constructing proxies of protected attributes within a model. Without careful measurement, analysis, and mitigation of these potential biases, it is not prudent to deploy AI models in the wild. These biases can be reduced via pre-, in-, or post-processing techniques-- intervening before, during, or after model training is complete.

Examples of pre-processing techniques include upweighting underrepresented populations in the training data, or learning representations of data that prevent AI/ML models from latching onto class membership, e.g. via debiasing word embeddings. If done correctly, these can prevent AI models from amplifying bias present within the data. Examples of in-processing techniques include adding terms to a model's objective function or imposing a fairness constraint during training; these methods can be extremely effective at creating fair outcomes, but only work with a well-defined consensus on the appropriate way to measure fairness for the given scenario, and may also create an accuracy-fairness tradeoff where the fairness objective is directly competing with the underlying model accuracy. Lastly, post-processing techniques might

[13] "Apple Card Investigated After Gender Discrimination Complaints ...." 10 Nov. 2019, https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html. Accessed 19 May. 2021.
[14] "Feature-Wise Bias Amplification." https://arxiv.org/abs/1812.08999. Accessed 19 May. 2021.

include setting unique classification thresholds for protected classes to balance fairness rates between groups; these techniques are rarely suitable for financial applications where disparate treatment of individuals based on protected attributes are prohibited.

These mitigation strategies, while useful, should never be blindly applied. Instead, careful consideration of the root cause of the bias by understanding the disparity in feature-level drivers of a model's decisions across groups of the population can elucidate what is going wrong, and guide the choice of the mitigation strategy.

13. To what extent do model risk management principles and practices aid or inhibit evaluations of AI-based credit determination approaches for compliance with fair lending laws?

MRM principles are core to ensuring AI/ML models are compliant with fair lending laws, but there are gaps in their practical usage. As an example, MRM principles of conceptual soundness can help assess whether model biases are justified or not (i.e. whether a model is using a reasonable feature which is simply correlated with a protected attribute, rather than directly using a proxy of this attribute).

However, in order to fully harness rigorous MRM practices for fairness, there needs to be a consensus on appropriate measures of fairness and how to quantify fairness for specific financial applications. In addition, just as model transparency/explainability is critical in justifying a model's output, root cause analysis to understand the underlying sources of bias is imperative to declare a model "fair". This can be accomplished via model transparency. By explaining why a model's predictions differ relative to pertinent comparison groups ("Why was Jane denied a loan relative to all approved men?"), and by digging into the different inputs that drive model decisions for members of different protected groups, the true source of the bias can be identified. Then, the choice of mitigation technique can be an informed decision.

Explainability can guide root cause analysis for fairness through model development and validation, underpinning both Takeaways 1 and 5.

15. The Equal Credit Opportunity Act (ECOA), which is implemented by Regulation B, requires creditors to notify an applicant of the principal reasons for taking adverse action for credit or to provide an applicant a disclosure of the right to request those reasons. What approaches can be used to identify the reasons for taking adverse action on a credit application, when AI is employed? Does Regulation B provide

sufficient clarity for the statement of reasons for adverse action when AI is used? If not, please describe in detail any opportunities for clarity.

To be compliant with ECOA/Regulation B, it is clear that FIs that choose to use AI must justify a model's decisions at a local/individual level to provide these "reason codes". Yet as we state above, the exact model outcome that is being explained can have vast implications on the reason code itself; explaining a classification decision ("Why was Jane denied a loan?") vs. a score decision ("Why was Jane given a risk score of 0.4?") can yield entirely different results. Thus, any explainability technique used to provide adverse action notices must be flexible in terms of these output types; Shapley value based attribution methods[15, 16], in particular, can break down many model outputs into its constituent features. However, different approximations of Shapley values (e.g. SHAP and QII) can yield dramatically and meaningfully different results for varying output types (e.g. classification vs. model scores). Thus, FIs require greater regulatory clarity on what output types are appropriate to explain and what techniques are best suited for these analyses.

It is also important to ensure that such customer-facing explanations meet a minimum standard of accuracy, which may well differ by materiality.

Nuances of adverse action codes include the output type that is being explained as well as the comparison group of the explanation; FIs would benefit from greater clarity on what output types to explain and which methods are well-suited to this.

17. To the extent not already discussed, please identify any benefits or risks to financial institutions' customers or prospective customers from the use of AI by those financial institutions. Please provide any suggestions on how to maximize benefits or address any identified risks.

AI can transform the way that financial institutions interact with customers, potentially enabling those that did not benefit from traditional models. But AI also can distance customers from grasping why a decision was made, leaving them unsatisfied and confused. If FIs thoroughly vet models for their *quality* (including transparency, stability, fairness and beyond) and learn effective ways of communicating this to their customers, only then will customers trust decisions made by AI.

---

[15] "A Unified Approach to Interpreting Model Predictions." 22 May. 2017, https://arxiv.org/abs/1705.07874. Accessed 19 May. 2021.
[16] "Algorithmic Transparency via Quantitative Input Influence:." https://www.andrew.cmu.edu/user/danupam/datta-sen-zick-oakland16.pdf. Accessed 19 May. 2021.