July 1, 2021

Board of Governors of the Federal Reserve System, Docket No. OP–1743
Consumer Financial Protection Bureau, Docket No. CFPB–2021– 0004
Federal Deposit Insurance Corporation, RIN 3064–ZA24
National Credit Union Administration, Docket No. NCUA –2021– 0023
Office of the Comptroller of the Currency, Docket ID OCC– 2020–0049
(Collectively, the "Agencies")

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

To: The Agencies

We the undersigned civil rights, consumer, technology, and other advocacy organizations are writing in response to the Agencies' March 31, 2021 Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning (the "RFI").[1]

We applaud the Agencies for seeking input on the critically-important topic of artificial intelligence ("AI") and machine learning ("ML"). Our organizations believe that the responses below will help inform the Agencies' policies and positions on the use of AI or ML by financial institutions.[2]

EXECUTIVE SUMMARY

In the RFI, the Agencies stated that they are seeking input to understand four broad areas:

- The use of AI by financial institutions;
- Appropriate governance, risk management, and controls over AI;
- Any challenges in developing, adopting, and managing AI; and
- Whether any clarifications from the Agencies would be helpful.

As the Agencies undertake this important challenge, our organizations recommend the following:

[1] Agencies, Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning, 86 Fed. Reg. 16837 (Mar. 31, 2021).

[2] Note on the language used in this response: There is no universal agreement on definitions for key terms such as "artificial intelligence," "race and ethnicity," and "fairness." We intend in all cases to be inclusive, rather than exclusive, and in no case to diminish the significance of the viewpoint of any person or to injure a person or group through our terminology. For the purposes of this response, we define "artificial intelligence" broadly to include a range of technologies and standardized practices, especially those that rely on machine learning or statistical theory. We use the following language with respect to race and ethnicity: Black, Latino, Asian American, and White. Instead of "fair" or "responsible" AI systems, we generally use the term "non-discriminatory" to refer to AI systems that do not disparately treat or impact people on a prohibited basis, and "equitable" to mean AI systems that promote equitable outcomes, particularly those that address historical discrimination. Finally, the term "bias" has several meanings depending on the context, so, to the extent possible, we have tried to clarify whether we mean racial bias, model bias, or other forms of bias.

- Non-discrimination and Equity: The Agencies should take the steps needed to ensure non-discriminatory and equitable outcomes for all who participate in the financial services market. Most importantly, the Agencies should define "model risk" to include the risk of discriminatory or inequitable outcomes for consumers, rather than just the risk of financial loss to a financial institution.[3] That is, the analysis of fair lending risk and equity should be integrated into all AI discussions and not treated as an afterthought. Throughout this document, we have demonstrated how the evaluation of fair lending risk and equity is an integral part of evaluating every aspect of AI risk.

- Action Plan: After review of the RFI responses, the Agencies should immediately issue a detailed Action Plan, which may include plans for a white paper regarding a proposed framework for the regulation of AI in financial services, a policy statement reminding financial institutions of their responsibilities under fair lending and other consumer protection laws, a policy statement describing the Agencies' expectations for financial institutions with respect to AI, a policy statement regarding the Agencies' methodologies for evaluating AI systems, a proposed regulation (including under the CFPB's UDAAP authority), and/or examination procedures.

- Robust Supervision and Enforcement/Accountability: The Agencies should conduct in-depth reviews of financial institutions' use of AI, including assessing compliance with fair lending laws.
    - Consistent with the Uniform Interagency Consumer Compliance Rating System[4] and the Model Risk Management Guidance,[5] the Agencies should ensure that financial institutions have appropriate Compliance Management Systems that effectively identify and control risks related to AI systems, including the risk of discriminatory or inequitable outcomes for consumers. The Compliance Management System should comprehensively cover the roles of board and senior management, policies and procedures, training, monitoring, and consumer complaint resolution. The extent and sophistication of the financial institution's Compliance Management System should align with the extent, sophistication, and risk associated with the financial institution's usage of the AI system, including the risk that the AI system could amplify historical patterns of discrimination in financial services.
    - Where a financial institution's use of AI indicates weaknesses in their Compliance Management System or violations of law, the Agencies should use all of the tools in their toolbelt to quickly address and prevent consumer harm, including issuing Matters Requiring Attention; entering into a non-public enforcement action, such as a

[3] *See* Federal Reserve Board and OCC, Supervisory Guidance on Model Risk Management, SR 11-7 at 3 (Apr. 4, 2011) ("Model Risk Management Guidance") (defining "model risk" to focus on the financial institution rather than the consumer by stating that "[m]odel risk can lead to financial loss, poor business and strategic decision making, or damage to a bank's reputation").

[4] FFIEC, Uniform Interagency Consumer Compliance Rating System at 21-22 (Nov. 7, 2016) (stating that for purposes of a financial institution's consumer compliance rating, examiners will assess the financial institution's Compliance Management System based on the board and management oversight as well as the compliance program, which includes policies and procedures, training, monitoring, and complaint resolution). *See also* CFPB Bulletin 2020-01, Responsible Business Conduct: Self-Assessing, Self-Reporting, Remediating, and Cooperating (Mar. 6, 2020).

[5] Model Risk Management Guidance at 16 (stating that "[d]eveloping and maintaining strong governance, policies, and controls over the model risk management framework is fundamentally important to its effectiveness").

Memorandum of Understanding; referring a pattern or practice of discrimination to the U.S. Department of Justice; or entering into a public enforcement action. The Agencies have already provided clear guidance (e.g., the Uniform Consumer Compliance Rating System) that financial institutions must appropriately identify, monitor, and address compliance risks, and the Agencies should not hesitate to act within the scope of their authority.

- Moreover, any new policies or initiatives related to AI should clearly state that the Agencies will hold financial institutions accountable for Compliance Management System weaknesses or violations of law.
- When possible, the Agencies should explain to the public the risks that they have observed and the actions taken in order to bolster the public's trust in robust oversight, and provide clear examples to guide the industry.

- Actionable Policies: Existing civil rights laws and policies provide a framework for the Agencies to analyze fair lending risk in AI and to engage in supervisory or enforcement actions, where appropriate. That said, the Agencies can be more effective in ensuring consistent and effective compliance by setting clear and robust regulatory expectations regarding testing and ensuring models are non-discriminatory and equitable. The Agencies have been in learning mode for some time, which may have put the U.S. behind in advancing non-discriminatory and equitable technology in financial services. To retain our competitive edge in the global society, the U.S. federal financial regulators should move quickly to issue actionable policy statements that clearly state their commitment to consumer protection and civil rights laws, including fair lending laws; insight into their supervisory expectations and methods; and useful guardrails and best practices. The time to act is now as the use of AI proliferates in every aspect of consumer financial services and has the potential for far-reaching adverse impacts for consumers of color and other protected groups. More specifically, the Agencies can be more effective in ensuring robust and consistent compliance by moving quickly to issue a clear policy statement on AI that:

  1. Defines "model risk" to include the risk of discriminatory or inequitable outcomes;
  2. Describes the risks that financial institutions should be aware of and control for;
  3. Sets clear standards for a financial institution's fair lending risk assessments, including:
     a. Discrimination testing and evaluation throughout the AI/ML model's conception, design, implementation, and use; and
     b. Information that must be detailed in the documentation of the financial institution's fair lending risk assessment, including:
        (i) What testing has been conducted and less discriminatory alternatives have been considered;
        (ii) In-depth information regarding the data that was used to train the model, measures taken to ensure the data was representative and accurate, and the attributes used in the model and its target outcomes; and
        (iii) Documentation on adverse action notices detailing the mechanism by which the adverse action notices are created and showing that the mechanism

provide adverse action notices that reliably produce consistent and specific reasons that consumers can understand and respond to, as appropriate;

4. Clarifies that the financial institution's fair lending risk assessment should be conducted by independent actors within the institution or a third party;[6]
5. Explains the metrics and methods that the Agencies will use to evaluate compliance with fair lending laws;
6. Sets documentation and archiving requirements sufficient to ensure that financial institutions maintain the data, code, and information necessary for Agencies to review their AI/ML systems;
7. Sets explainability standards sufficient to enable the Agencies, advocates, consumers, independent auditors, and other key stakeholders to understand the decisions and outcomes generated by AI systems;
8. States that the Agencies will test for fair lending risk consistent with fair lending laws and policies, including by:
    a. Testing for disparate impact and less discriminatory alternatives;
    b. Ensuring that the training data is representative and accurate;
    c. Ensuring that the model measures lawful and meaningful attributes and seeks to predict valid target outcomes; and
    d. Ensuring that the technology is interpretable and its decision-making is sufficiently explainable to comply with fair lending laws;
9. To the maximum extent possible, ensures public access to detailed information about financial institutions' use of AI/ML and assessments of those models as well as the Agencies' reviews; and
10. Provides examples of best practices that financial institutions can use to mitigate fair lending risk.

- Public Research: The Agencies should encourage and support public research that analyzes the efficacy of specific uses of AI in financial services and the impact of AI in financial services for consumers of color and other protected classes. For example, the Agencies should encourage the CFPB and the Federal Housing Finance Agency to release more de-personalized loan-level data from the National Survey of Mortgage Originations and the National Mortgage Database so researchers, advocacy groups, and the public can study potential discriminatory and inequitable outcomes in the financial services market, especially as they relate to the use of AI.
    - For example, a research partnership could be formed between the National Institute of Standards and Technology ("NIST"), civil rights organizations, consumer protection groups, non-profit research agencies, and financial institutions that rely on AI to evaluate how AI or machine learning models affect fair lending.[7]
    - The Agencies may also work with the National Science Foundation to ensure that a portion of the considerable allocations for research on AI focus on the implications of using AI in financial services.

[6] This approach is consistent with the Model Risk Management Guidance, which states: "Validation involves a degree of independence from model development and use. Generally, validation should be done by people who are not responsible for development or use and do not have a stake in whether a model is determined to be valid." Model Risk Management Guidance at 9.

[7] *See, e.g.*, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software, NIST (Dec. 19, 2019).

- The Agencies should find ways to include civil rights organizations with experience in working on fair lending issues on research projects.

- Specialized Fair Lending and AI Staff: The Agencies should immediately begin hiring staff with specialized skills that can provide guidance to financial institutions on assessing the impact of AI systems and that can review those assessments, particularly with respect to fair lending risks.
  - The Agencies should ensure that a financial institution's use of AI is reviewed by agency staff that specialize in AI and fair lending risks.
  - In addition, the Agencies should remind examination teams of the requirement under the Equal Credit Opportunity Act (ECOA) to examine for discriminatory impacts and less discriminatory alternatives, and to refer a matter to the U.S. Department of Justice if the agency has reason to believe that a creditor has engaged in a pattern or practice of discrimination.
  - Finally, given that AI models cannot be fully evaluated without considering the high risk of far-reaching and repeatable disparate impact, each Agency's fair lending and AI teams should be included in all meetings related to modeling efforts, including supervisory and enforcement issues, policy statements, and rulemakings.

- Fair Lending Training for All AI Stakeholders: The Agencies should ensure that all AI stakeholders—including regulators, financial institutions, and tech companies—receive regular fair lending and racial equity training. Trained professionals are better able to identify and recognize issues that may raise red flags. They are also better able to design AI systems that generate non-discriminatory and equitable outcomes. The more stakeholders in the field are educated about fair lending and equity issues, the more likely they are to create tools that expand opportunities for all consumers. Given the ever-evolving nature of AI, the training should be updated and provided on a periodic basis.

- Diversity, Equity, and Inclusion: The Agencies should ensure agency staff working on AI issues reflect diversity, including diversity based on race and national origin. In addition, the Agencies should encourage financial institutions to engage diverse staff for the AI development and design teams. Increasing the diversity of the regulatory and industry staff engaged in AI issues will lead to better outcomes for consumers. Research has shown that diverse teams are more innovative and productive[8] and that companies with more diversity are more profitable.[9] Moreover, people with diverse backgrounds and experiences bring unique and important perspectives to understanding how data impacts

[8] *See, e.g.*, John Rampton, Why You Need Diversity on Your Team, and 8 Ways to Build It, Entrepreneur (Sept. 6, 2019).

[9] *See, e.g.*, David Rock and Heidi Grant, Why Diverse Teams Are Smarter, Harvard Business Review (Nov. 4, 2016) (reporting that companies in the top quartile for ethnic and racial diversity in management were 35% more likely to have financial returns above their industry mean, and those in the top quartile for gender diversity were 15% more likely to have returns above the industry mean).

different segments of the market.[10] In several instances, it has been people of color who were able to identify potentially discriminatory AI systems.[11]

- Transparency: The Agencies should prioritize transparency as they develop their understanding of the issues and proposed responses. First, the Agencies should strive to share their data, models, decisions, and proposed solutions so that all of the key stakeholders can stay apprised of and comment on the potential impact of proposed Agency actions on the national consumer financial market. Second, the Agencies should require financial institutions to share with the public as much information as possible regarding their AI systems and assessments of those systems to enable researchers and those impacted to evaluate the efficacy and impact of the systems.

- Engagement: The Agencies should stay engaged with a diverse group of key stakeholders, including civil rights organizations, consumer advocates, and impacted communities in order to receive ongoing input and feedback on these important decisions. The proposed solutions to AI risks are likely to have significant implications for borrowers and communities of color as well as other vulnerable communities, such as individuals with disabilities, families, and Limited English Proficiency borrowers. The Agencies should regularly engage with these communities and seek solutions that treat all borrowers and communities equitably.

[10] *See, e.g.*, Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, *in* Conference on Fairness, Accountability, and Transparency 33, 39 (2020) (stressing the importance of "standpoint diversity" as algorithm development implicitly encodes developer assumptions of which they may not be aware). *See also* Model Risk Management Guidance at 4 (stating that "[a] guiding principle for managing model risk is 'effective challenge' of models, that is, critical analysis by objective, informed parties who can identify model limitations and assumptions and produce appropriate changes").

[11] *See, e.g.*, Steve Lohr, Facial Recognition is Accurate, if You're a White Guy, New York Times (Feb. 9, 2018) (explaining how Joy Buolamwini, a Black computer scientist, discovered that facial recognition worked well for her White friends but not for her).

BACKGROUND AND GLOBAL RECOMMENDATIONS

AI Has the Potential to Perpetuate, Amplify, and Accelerate Historical Patterns of Discrimination

For much of America's history, communities of color were systematically excluded from economic opportunities through explicit government policy decisions that inculcated an inappropriate and unfounded association between race and risk into the nation's housing and financial markets. In particular, the New Deal's federal Home Owners Loan Corporation ("HOLC")[12] developed one of the most harmful policy decisions in the housing market by creating a mapping system that included race as a fundamental factor in determining the desirability of neighborhoods.[13] Notably, the data used to create the maps were not just collected randomly, but rather were based on the racist views of the leading real estate professionals at the time. Based on this feedback, the HOLC coded communities of color as "hazardous." These areas were designated by red shading on the Residential Security Survey maps created by the HOLC and were assigned a lower value.[14] This approach systematized the link between race and risk and institutionalized "redlining," which refers to restricting access to credit in communities of color.

Later, the Federal Housing Administration adopted these maps as the basis for its mortgage insurance underwriting decisions. Thus, the maps not only reflected the race-based views of the nation's housing industry leaders at the time, but were also used to amplify and codify these views throughout the housing system. These discriminatory policies and several others created distinct advantages for White families, leading to massive wealth, homeownership, and credit gaps between White families and families of color that persist today.[15]

Right now, America is at a similar crossroads in determining whether or how to develop equitable AI systems that serve and uplift the whole of the national financial services market, or systems that perpetuate, amplify, and even accelerate existing discriminatory patterns. The time to act is now as the use of AI in financial services proliferates in every aspect of consumer financial services and has the potential for far-reaching adverse impacts for borrowers of color and other protected groups that could overshadow even the devastation caused by the HOLC, the FHA, and other entities that perpetuated discriminatory practices. Government, industry, and

---

[12] The Home Owners' Loan Act of 1933 established the HOLC as an emergency agency under the Federal Home Loan Bank Board. 12 U.S.C. § 1461 *et seq*.

[13] *See* Lisa Rice, "The Fair Housing Act: A Tool for Expanding Access to Quality Credit," The Fight for Fair Housing: Causes, Consequences, and Future Implications of the 1968 Federal Fair Housing Act (Gregory Squires, 1st ed. 2017) (providing a detailed explanation of how federal race-based housing and credit policies promoted inequality). *See also*, K. Steven Brown et al., Confronting Structural Racism in Research and Policy Analysis, The Urban Institute (Feb. 2019); Richard Rothstein, The Color of Law: A Forgotten History of How Our Government Segregated America (2017).

[14] *See* University of Richmond, Virginia Tech, University of Maryland, and Johns Hopkins University, Mapping Inequality (documenting the maps and area descriptions created by the HOLC between 1935 and 1940).

[15] *See* Neil Bhutta et al., Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances, FEDS Notes, Board of Governors of the Federal Reserve System (Sept. 2020); Heather Long and Andrew Van Dam, The Black-White Economic Divide Is as Wide as It Was in 1968, Washington Post (June 4, 2020); Bruce Mitchell and Juan Franco, HOLC "Redlining" Maps: The Persistent Structure of Segregation and Economic Inequality, National Community Reinvestment Coalition (Feb. 2018).

advocacy groups should work together to ensure that AI systems support non-discriminatory and equitable housing and finance markets. This is simply the right thing to do, and will benefit individual consumers and our whole society.

Existing Civil Rights Laws and Policies Provide a Framework for Analyzing Fair Lending Risk in AI and Engaging in Appropriate Supervisory or Enforcement Actions

Two primary federal anti-discrimination laws—ECOA and the Fair Housing Act (collectively, the "fair lending laws")—prohibit institutions from discriminating in lending and housing on the basis of characteristics such as race, national origin, religion, and sex.[16] ECOA applies to nearly all lending, including lending to businesses. The Fair Housing Act applies to housing discrimination, including discrimination in mortgage lending and other residential real estate-related transactions.[17] In addition, the Agencies have issued several policies that provide a framework for effective supervision and enforcement as well as guidance to financial institutions, including the Interagency Fair Lending Examination Procedures, the Model Risk Management Guidance, the Uniform Interagency Consumer Compliance Rating System, and the Bulletin on Responsible Business Conduct.[18]

The fair lending laws prohibit policies and practices when there is evidence of intentional discrimination, known as "disparate treatment," as well as when—even without evidence of discriminatory intent—there is evidence of a discriminatory effect called "disparate impact." Disparate treatment occurs when an entity explicitly or intentionally treats people differently based on protected characteristics, such as race, national origin, or sex. In contrast, disparate impact focuses on outcomes. Generally, unlawful disparate impact occurs when a (1) facially neutral policy or practice disproportionately adversely impacts members of protected classes, and either (2) the policy or practice does not advance a legitimate interest, or (3) is not the least discriminatory means to advance that interest.[19] These frameworks translate well to supervisory reviews of consumer financial services models, including AI/ML models, although more guidance would be helpful to ensure robust, consistent, and effective application.

The methodologies that regulators and financial institutions use for fair lending testing models can vary, but as a general matter the most effective systems are designed to align with regulatory expectations and traditional principles gleaned from anti-discrimination jurisprudence. These systems often include: (1) ensuring that models do not include protected characteristics or close proxies for protected characteristics, for example as variables or segmentations; and (2) assessing whether facially-neutral models are likely to disproportionately lead to negative outcomes for a

---

[16] 15 U.S.C. § 1691(a); 12 C.F.R. § 1002.2(z); 42 U.S.C. § 3605. *See also* Civil Rights Act of 1866, 42 U.S.C. §§ 1981, 1982.

[17] Other laws, like the Agencies' UDA(A)P authorities, can also be applied to prevent discrimination in consumer financial services, although to date the Agencies have not leveraged their UDA(A)P authorities to combat discrimination. *See* Stephen Hayes and Kali Schellenberg, Discrimination is "Unfair": Interpreting UDA(A)P to Prohibit Discrimination, Student Borrower Protection Center (Apr. 2021).

[18] FFIEC, Revised FFIEC Fair Lending Examination Procedures and Use of Specialized Examination Techniques, (Aug. 4, 2009); Model Risk Management Guidance at 3; FFIEC, Uniform Interagency Consumer Compliance Rating System (Nov. 7, 2016); CFPB Bulletin 2020-01, Responsible Business Conduct: Self-Assessing, Self-Reporting, Remediating, and Cooperating (Mar. 6, 2020).

[19] *See, e.g.,* 12 C.F.R. Part 1002, Supp. I, ¶ 6(a)-2 (ECOA articulation); 24 C.F.R. § 100.500(c)(1) (FHA articulation); 42 U.S.C. § 2000e-2(k) (Title VII articulation).

protected class, and if such negative impacts exist, ensuring the models serve legitimate business needs and evaluating whether changes to the models would result in less of a disparate impact while maintaining model performance.[20]

Moreover, in robust Compliance Management Systems, financial institutions will augment these (and other) quantitative statistical tests with more holistic compliance controls: ensuring effective board and management oversight; ensuring robust model governance; reviewing policies and procedures within which models operate, including credit policies, overlays, exclusions, overrides and the like; assessing areas of discretion to ensure that the potential for judgmental bias is mitigated; providing fair lending training for relevant staff, including modelers; ensuring teams have diverse backgrounds and are empowered to identify and remedy issues; ensuring effective monitoring, including independent compliance auditing; and ensuring effective consumer complaint resolution processes.

In short, existing civil rights laws and supervisory policies provide a framework for the Agencies to analyze fair lending risk in AI and to engage in supervisory or enforcement actions, where appropriate. That said, the Agencies should ensure consistent and effective compliance by, among other things, clarifying certain ambiguities, setting robust regulatory expectations regarding testing, and ensuring models are non-discriminatory and equitable, as discussed below.

Recent Agency Actions Have Failed to Show a Clear Commitment to Civil Rights

In recent policy activities, the Agencies have sometimes failed to give civil rights and consumer protections the proper weight. For example, the CFPB's recent actions have seemed to favor regulatory flexibility while lacking clear guidance for robust civil rights and consumer protection, particularly for borrowers of color.

- Upstart No Action Letters: Between 2017 and 2020 the CFPB issued two No Action Letters ("NAL") (and an Agency "update") to Upstart Network ("Upstart") to facilitate innovation.[21] We are concerned, however, that it did so without fully accounting for certain aspects of the company's model that have long been recognized as having a disparate impact on borrowers of color. Upstart is a lending platform that relies on ML-based AI models and non-traditional applicant data to underwrite and price consumer loans. The CFPB's NAL program is meant to foster innovation by issuing a non-binding letter expressing that the CFPB has no present intention to bring enforcement or supervisory actions against the institution related to the product or service in question.[22] In this case, the CFPB relied on the representations by the company for its determination and did not replicate the company's analysis. Subsequently, the NAACP Legal Defense

[20] *See* Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model, Initial Report of the Independent Monitor, 7 (Apr. 14, 2021); Nicholas Schmidt and Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination, Consumer Finance Law Quarterly Report, Vol. 73(2) 130, 141–142 (2019); David Skanderson and Dubravka Ritter, Fair Lending Analysis of Credit Cards, Federal Reserve Bank of Philadelphia, 38–40 (2014).
[21] CFPB, Letter Response to 2017 NAL Request (Sept. 14, 2017); Patrice Alexander Ficklin and Paul Watkins, An Update on Credit Access and the Bureau's First No-Action Letter (Aug. 6, 2019); CFPB, Letter Response to 2020 NAL Request (Nov. 30, 2020).
[22] CFPB, Policy on No-Action Letters; Information Collection, 81 Fed. Reg. 8686, 8686 (Feb. 22, 2016).

Fund ("LDF") and the Student Borrower Protection Center ("SBPC") raised concerns that Upstart's use of certain education data posed a high risk of disparate impact for borrowers of color. As a result, Upstart, LDF, and the SBPC have engaged the services of a civil rights law firm to act as an independent fair lending monitor to evaluate and make recommendations regarding the fair lending implications of Upstart's lending platform.[23] At this time, Upstart has already implemented certain modifications to address disparate impact concerns in its underwriting and pricing models, including establishing a "normalization" process for Minority Serving Institutions and eliminating the use of average incoming SAT and ACT scores to group education institutions. Future periodic reports by the independent fair lending monitor will address fair lending tests conducted and recommendations specific to Upstart's model following the adoption of these changes. In the meantime, the use of Upstart's lending platform has proliferated from use by one FDIC-regulated community bank (Cross River Bank) to use by several financial institutions through the institutions' web portals. The use of Upstart's model is not transparent to the average consumer.

- Innovation Spotlight regarding Adverse Action Notices: In July of 2020, the CFPB released a blog post regarding ECOA adverse action notices that use AI/ML models. The blog post emphasized the "flexibility" of the regulation for AI decisions based on complex interrelationships. However, the post did not show a strong commitment to ensuring that AI providers and users adhere to the letter and spirit of ECOA, which was meant to ensure that consumers could understand the credit denials that impact their lives, discern whether the reason provided was the real reason, and determine that there was no unlawful discriminatory reason.[24] The CFPB stated that although a creditor must provide the specific reasons for an adverse action, the Official Interpretation to Regulation B, which implements ECOA, provides that a creditor need not describe how or why a disclosed factor adversely affected an application,[25] or, for credit scoring systems, how the factor relates to creditworthiness.[26] The CFPB emphasized that "a creditor may disclose a reason for a denial even if the relationship of that disclosed factor to predicting creditworthiness may be unclear to the applicant." The CFPB went on to highlight a variety of tools that the industry can use as "safe harbors" for innovation.

- Tech Sprint regarding Adverse Action Notices: In October 2020, the CFPB hosted a "tech sprint" where organizations could develop innovative ways to notify consumers of adverse credit actions.[27] While this represented a creative approach to the issue, to date, there has been little to indicate the extent to which the CFPB plans to hold providers and users of AI responsible for providing meaningful adverse action notices as required by ECOA.

---

[23] *See* Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model, Initial Report of the Independent Monitor, (Apr. 14, 2021). *See also* Democratic Members of the Senate Committee on Banking, Housing, and Urban Affairs, Use of Educational Data to Make Credit Determinations (July 2020) (raising concerns about Upstart's use of non-individualized cohort-level education data).
[24] Patrice Alexander Ficklin, Tom Pahl, and Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices When Using AI/ML Models, CFPB Blog (July 7, 2020).
[25] 12 C.F.R. Part 1002, Supp. I, ¶ 9(b)(2)-3.
[26] *Id*. at ¶ 9(b)(2)-4.
[27] CFPB, Tech Sprint on Electronic Disclosures of Adverse Action Notices (Oct. 5-9, 2020).

Similarly, the other Agencies have not issued any policy statements or brought any enforcement actions signaling a strong commitment to holding providers and users of AI systems responsible for compliance with fair lending and other consumer protection laws.[28] By comparison, the Federal Trade Commission ("FTC") recently issued a strong warning to the industry regarding AI to "keep in mind that if you don't hold yourself accountable, the FTC may do it for you." We hope this RFI signals a change in the interest and commitment of the Agencies.[29]

The U.S. Federal Financial Regulators Are Behind in Advancing Non-discrimination and Equity in AI, but There Are Examples of Useful Starting Points for a Robust Regulatory Framework

In some respects, the U.S. federal financial regulators are behind in advancing non-discriminatory and equitable technology for financial services. If we want to retain our competitive edge in the global society, we should hasten to minimize harm from existing technologies and take the necessary steps to ensure all AI systems generate non-discriminatory and equitable outcomes. Moreover, the transition from incumbent models to AI-based systems presents an important opportunity to address what is wrong in the status quo—baked-in disparate impact and a limited view of the recourse of consumers who are harmed by current practices—and to rethink appropriate guardrails to promote a safe, fair, and inclusive financial sector. The Agencies have an opportunity to rethink comprehensively how they regulate key decisions that determine who has access to financial services and on what terms.

As the Agencies consider their approach to the use of AI in financial services, the European Union's newly-released proposed regulation for AI ("EU Proposed Regulation") may be a useful example of how to define "model risk" to include the risk of discriminatory or inequitable outcomes for consumers (rather than just financial loss for financial institutions).[30] Notably, the EU's risk-focused framework recognizes that AI systems that impact the evaluation of creditworthiness risks violating widely recognized anti-discrimination protections, and should be strictly regulated.[31] Importantly, the EU made this determination based on explicit recognition of (i) the importance of credit evaluations to fully participate in society or improve one's standard of living and (ii) the high risk of discrimination. The preamble to the proposed regulation states:

> Another area in which the use of AI systems deserves special consideration is the access to and enjoyment of certain essential private and public services and benefits **necessary for people to fully participate in society or to improve one's standard of living**. In

[28] Elisa Jillson, Aiming for Truth, Fairness, and Equity in Your Company's Use of AI, Federal Trade Commission (Apr. 19, 2021).
[29] We note that ensuring that AI/ML models are properly reviewed for compliance with civil rights laws would be consistent with the Administration's Executive Order on Racial Equity.
[30] European Commission, Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (also known as the "Artificial Intelligence Act") (Apr. 21, 2021). Notably, the Proposed Regulation would apply to both providers and users of AI systems, including those that are located outside of the EU if the output is used in the EU. Thus, although the EU Proposed Regulation highlights a gap in U.S. oversight, it may ultimately reduce the costs and resistance to compliance with any new policies or regulations promulgated by U.S. federal financial regulators as a significant set of American firms may be already complying with the EU's framework.
[31] *Id*. at Annex III.

> particular, AI systems used to evaluate the credit score or creditworthiness of natural persons should be classified as high-risk AI systems, since they determine those persons' access to financial resources or essential services such as housing, electricity, and telecommunication services. AI systems used for this purpose may lead to discrimination of persons or groups and **perpetuate historical patterns of discrimination,** for example based on racial or ethnic origins, disabilities, age, sexual orientation, or **create new forms of discriminatory impacts**.[32]

Thus, the EU recognizes that AI systems that evaluate creditworthiness should be held to a high standard given the far-reaching impact on consumers' life options and the high risk of discrimination.

The EU and Others Provide a Useful Starting Point for a Robust Regulatory Framework

The EU's Proposed Regulation and other laws and policies provide a useful starting point for building off existing U.S. anti-discrimination law and ensuring a robust regulatory framework for high-risk AI systems. For example, the EU Proposed Regulation would require providers to implement controls related to the following:

- Data governance,
- Transparency,
- Human oversight,
- Risk and quality management systems,
- Security, and
- Post-deployment monitoring.[33]

Moreover, a provider of a high-risk AI system would need to conduct a conformity assessment and certify the system's conformity with the regulation *before* the system is released to the market.[34] Finally, providers of AI systems must ensure that natural persons are informed that they are interacting with an AI system.[35] Penalties by regulators for non-compliance would be as high as 6% of the entity's total global earnings (before costs).[36]

Regulators in the U.S. should further improve on the EU proposal by ensuring transparency and Agency review of provider self assessments.[37] The Agencies should require that financial institutions conduct fair lending risk assessments that detail how their AI/ML systems were trained and tested in their design, implementation, and use, including for disparate impact; detail the training data used, the attributes used in the model and the target outcomes; and assess the outcomes and impact of their models. The Agencies should require that financial institutions

---

[32] *Id*. at Recital 37 (emphasis added).
[33] *Id*. at Titles III and VIII.
[34] *Id*. at Title III, Ch. 3 and 5.
[35] *Id*. at Title IV, Art. 52.
[36] *Id*. at Title X, Art. 71.
[37] *See, e.g.*, Mark MacCarthy and Kenneth Propp, Machines Learn That Brussels Writes the Rules: The EU's New AI Regulation, Brookings Institution (May 4, 2021); Adam Satariano, Europe Proposes Strict Rules for Artificial Intelligence, NewYork Times (Apr. 21, 2021) (quoting an advocacy group that is critical of the Proposed Regulation's reliance on self assessments).

routinely provide their self assessments to the regulators for review, and also provide these self assessments to users and the public, to the maximum extent possible. Similar to the public availability of Home Mortgage Disclosure Act (HMDA) data, the public availability and transparency of self assessments (including disparate impact assessments) could help extend regulator resources, by facilitating independent external audits. This transparent approach would also facilitate exercise of private rights of action under existing consumer protection or civil rights laws, when appropriate, to ensure consumers are protected from discriminatory outcomes.

In addition to the EU Proposed Regulation, the Agencies may also find it instructive to review recent actions by legislators in New York and California. Legislators in New York introduced a bill that requires government agencies that seek to procure or use an AI/ML decision system to engage a neutral third party to conduct a civil rights assessment for public release and undergo a public hearing on the tool. The impact assessment includes "[a] detailed description of the automated decision system, its design, its training, its data, and its purpose;" a cost/benefit analysis; a risk assessment that includes "the risk that such automated decision system may result in or contribute to inaccurate, unfair, biased, or discriminatory decisions impacting individuals;" and a risk minimization plan. The bill also requires the development of usage policies, notice to individuals that an automated decision system was used, and the ability for individuals to contest its decision and obtain human review.[38]

In California, legislators introduced a similar bill requiring state agencies to minimize the discriminatory impacts of automated decision systems in state contracts. The bill provides that an application for a state contract is not considered complete until an applicant has described "any potential disparate impacts on the basis of characteristics identified in the Unruh Civil Rights Act (Section 51 of the Civil Code) from the proposed use of the automated decision system." Applications must also include "the extent to which members of the public have access to the results of the automated decision system, including an explanation for the basis of a resulting decision in terms understandable to a layperson, and are able to correct or object to its results, and where and how that information will be made available and any applicable procedures for initiating corrections or objections, as appropriate."[39]

Recommendations

Accordingly, our organizations recommend the following:

- Non-discrimination and Equity: The Agencies should take the steps needed to ensure non-discriminatory and equitable outcomes for all who participate in the financial services market. Most importantly, the Agencies should define "model risk" to include the risk of discriminatory or inequitable outcomes for consumers, rather than just the risk

[38] AB-A06042, Gen. Assemb., Reg. Sess. (N.Y. 2021-2022) ("NY State Digital Fairness Act"). *See also* Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, *in* Conference on Fairness, Accountability, and Transparency 33, 39 (2020); Senator Markey's Algorithmic Justice Online Privacy and Transparency Act; Personal Data Protection Commission ("PDPC") of Singapore, Singapore's Approach to AI Governance; PDPC of Singapore, Model AI Governance Framework, (2nd ed. 2020).

[39] AB-13, Gen. Assemb., Reg. Sess. (Cal. 2021-2022) ("California Automated Decision Systems Accountability Act").

of financial loss to a financial institution.[40] That is, the analysis of fair lending risk and equity should be integrated into all AI discussions and not treated as an afterthought. Throughout this document, we have demonstrated how the evaluation of fair lending risk and equity is an integral part of evaluating every aspect of AI risk.

- Action Plan: After review of the RFI responses, the Agencies should immediately issue a detailed Action Plan, which may include plans for a white paper regarding a proposed framework for the regulation of AI in financial services, a policy statement reminding financial institutions of their responsibilities under fair lending and other consumer protection laws, a policy statement describing the Agencies' expectations for financial institutions with respect to AI, a policy statement regarding the Agencies' methodologies for evaluating AI systems, a proposed regulation (including under the CFPB's UDAAP authority), and/or examination procedures.

- Robust Supervision and Enforcement/Accountability: The Agencies should conduct in-depth reviews of financial institutions' use of AI, including assessing compliance with fair lending laws.
  - Consistent with the Uniform Interagency Consumer Compliance Rating System[41] and the Model Risk Management Guidance,[42] the Agencies should ensure that financial institutions have appropriate Compliance Management Systems that effectively identify and control risks related to AI systems, including the risk of discriminatory or inequitable outcomes for consumers. The Compliance Management System should comprehensively cover the roles of board and senior management, policies and procedures, training, monitoring, and consumer complaint resolution. The extent and sophistication of the financial institution's Compliance Management System should align with the extent, sophistication, and risk associated with the financial institution's usage of the AI system, including the risk that the AI system could amplify historical patterns of discrimination in financial services.
  - Where a financial institution's use of AI indicates weaknesses in their Compliance Management System or violations of law, the Agencies should use all of the tools in their toolbelt to quickly address and prevent consumer harm, including issuing Matters Requiring Attention; entering into a non-public enforcement action, such as a Memorandum of Understanding; referring a pattern or practice of discrimination to the U.S. Department of Justice; or entering into a public enforcement action. The Agencies have already provided clear guidance (e.g., the Uniform Consumer Compliance Rating System) that financial institutions must appropriately identify,

[40] *See* Model Risk Management Guidance at 3 (defining "model risk" to focus on the financial institution rather than the consumer by stating that "[m]odel risk can lead to financial loss, poor business and strategic decision making, or damage to a bank's reputation").
[41] FFIEC, Uniform Interagency Consumer Compliance Rating System (Nov. 7, 2016) (stating that for purposes of a financial institution's consumer compliance rating, examiners will assess the financial institution's Compliance Management System based on the board and management oversight as well as the compliance program, which includes policies and procedures, training, monitoring, and complaint resolution). *See also* CFPB Bulletin 2020-01, Responsible Business Conduct: Self-Assessing, Self-Reporting, Remediating, and Cooperating (Mar. 6, 2020).
[42] Model Risk Management Guidance at 16 (stating that "[d]eveloping and maintaining strong governance, policies, and controls over the model risk management framework is fundamentally important to its effectiveness").

monitor, and address compliance risks, and the Agencies should not hesitate to act within the scope of their authority.
    - Moreover, any new policies or initiatives related to AI should clearly state that the Agencies will hold financial institutions accountable for Compliance Management System weaknesses or violations of law.
    - When possible, the Agencies should explain to the public the risks that they have observed and the actions taken in order to bolster the public's trust in robust oversight and provide clear examples to guide the industry.

- Actionable Policies: Existing civil rights laws and policies provide a framework for the Agencies to analyze fair lending risk in AI and to engage in supervisory or enforcement actions, where appropriate. That said, the Agencies can be more effective in ensuring consistent and effective compliance by setting clear and robust regulatory expectations regarding testing and ensuring models are non-discriminatory and equitable. The Agencies have been in learning mode for some time, which may have put the U.S. behind in advancing non-discriminatory and equitable technology in financial services. To retain our competitive edge in the global society, the U.S. federal financial regulators should move quickly to issue actionable policy statements that clearly state their commitment to consumer protection and civil rights laws, including fair lending laws; insight into their supervisory expectations and methods; and useful guardrails and best practices. The time to act is now as the use of AI proliferates in every aspect of consumer financial services and has the potential for far-reaching adverse impacts for consumers of color and other protected groups. More specifically, the Agencies can be more effective in ensuring robust and consistent compliance by moving quickly to issue a clear policy statement on AI that:

    1. Defines "model risk" to include the risk of discriminatory or inequitable outcomes;
    2. Describes the risks that financial institutions should be aware of and control for;
    3. Sets clear standards for a financial institution's fair lending risk assessments, including:
        a. Discrimination testing and evaluation throughout the AI/ML model's conception, design, implementation, and use; and
        b. Information that must be detailed in the documentation of the financial institution's fair lending risk assessment, including:
            (i) What testing has been conducted and less discriminatory alternatives have been considered;
            (ii) In-depth information regarding the data that was used to train the model, measures taken to ensure the data was representative and accurate, and the attributes used in the model and its target outcomes; and
            (iii) Documentation on adverse action notices detailing the mechanism by which the adverse action notices are created and showing that the mechanism provides adverse action notices that reliably produce consistent and specific reasons that consumers can understand and respond to, as appropriate;

4. Clarifies that the financial institution's fair lending risk assessment should be conducted by independent actors within the institution or a third party;[43]
5. Explains the metrics and methods that the Agencies will use to evaluate compliance with fair lending laws;
6. Sets documentation and archiving requirements sufficient to ensure that financial institutions maintain the data, code, and information necessary for Agencies to review their AI/ML systems;
7. Sets explainability standards sufficient to enable the Agencies, advocates, consumers, independent auditors, and other key stakeholders to understand the decisions and outcomes generated by AI systems;
8. States that the Agencies will test for fair lending risk consistent with fair lending laws and policies, including by:
    e. Testing for disparate impact and less discriminatory alternatives;
    f. Ensuring that the training data is representative and accurate;
    g. Ensuring that the model measures lawful and meaningful attributes and seeks to predict valid target outcomes; and
    h. Ensuring that the technology is interpretable and its decision-making is sufficiently explainable to comply with fair lending laws;
9. To the maximum extent possible, ensures public access to detailed information about financial institutions' use of AI/ML and assessments of those models as well as the Agencies' reviews; and
10. Provides examples of best practices that financial institutions can use to mitigate fair lending risk.

- Public Research: The Agencies should encourage and support public research that analyzes the efficacy of specific uses of AI in financial services and the impact of AI in financial services for consumers of color and other protected classes. For example, the Agencies should encourage the CFPB and the Federal Housing Finance Agency to release more de-personalized loan-level data from the National Survey of Mortgage Originations and the National Mortgage Database so researchers, advocacy groups, and the public can study potential discriminatory and inequitable outcomes in the financial services market, especially as they relate to the use of AI.
    - For example, a research partnership could be formed between the National Institute of Standards and Technology ("NIST"), civil rights organizations, consumer protection groups, non-profit research agencies, and financial institutions that rely on AI to evaluate how AI or machine learning models affect fair lending.[44]
    - The Agencies may also work with the National Science Foundation to ensure that a portion of the considerable allocations for research on AI focus on the implications of using AI in financial services.
    - The Agencies should find ways to include civil rights organizations with experience in working on fair lending issues on research projects.

[43] This approach is consistent with the Model Risk Management Guidance, which states: "Validation involves a degree of independence from model development and use. Generally, validation should be done by people who are not responsible for development or use and do not have a stake in whether a model is determined to be valid." Model Risk Management Guidance at 9.
[44] *See, e.g.*, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software, NIST (Dec. 19, 2019).

- Specialized Fair Lending and AI Staff: The Agencies should immediately begin hiring staff with specialized skills that can provide guidance to financial institutions on assessing the impact of AI systems and that can review those assessments, particularly with respect to fair lending risks.
    - The Agencies should ensure that a financial institution's use of AI is reviewed by agency staff that specialize in AI and fair lending risks.
    - In addition, the Agencies should remind examination teams of the requirement under ECOA to examine for discriminatory impacts and less discriminatory alternatives, and to refer a matter to the U.S. Department of Justice if the agency has reason to believe that a creditor has engaged in a pattern or practice of discrimination.
    - Finally, given that AI models cannot be fully evaluated without considering the high risk of far-reaching and repeatable disparate impact, each Agency's fair lending and AI teams should be included in all meetings related to modeling efforts, including supervisory and enforcement issues, policy statements, and rulemakings.

- Fair Lending Training for All AI Stakeholders: The Agencies should ensure that all AI stakeholders—including regulators, financial institutions, and tech companies—receive regular fair lending and racial equity training. Trained professionals are better able to identify and recognize issues that may raise red flags. They are also better able to design AI systems that generate non-discriminatory and equitable outcomes. The more stakeholders in the field are educated about fair lending and equity issues, the more likely they are to create tools that expand opportunities for all consumers. Given the ever-evolving nature of AI, the training should be updated and provided on a periodic basis.

- Diversity, Equity, and Inclusion: The Agencies should ensure agency staff working on AI issues reflect diversity, including diversity based on race and national origin. In addition, the Agencies should encourage financial institutions to engage diverse staff for the AI development and design teams. Increasing the diversity of the regulatory and industry staff engaged in AI issues will lead to better outcomes for consumers. Research has shown that diverse teams are more innovative and productive[45] and that companies with more diversity are more profitable.[46] Moreover, people with diverse backgrounds and experiences bring unique and important perspectives to understanding how data impacts

[45] *See, e.g.,* John Rampton, Why You Need Diversity on Your Team, and 8 Ways to Build It, Entrepreneur (Sept. 6, 2019).
[46] *See, e.g.*, David Rock and Heidi Grant, Why Diverse Teams Are Smarter, Harvard Business Review (Nov. 4, 2016) (reporting that companies in the top quartile for ethnic and racial diversity in management were 35% more likely to have financial returns above their industry mean, and those in the top quartile for gender diversity were 15% more likely to have returns above the industry mean).

different segments of the market.[47] In several instances, it has been people of color who were able to identify potentially discriminatory AI systems.[48]

- Transparency: The Agencies should prioritize transparency as they develop their understanding of the issues and proposed responses. First, the Agencies should strive to share their data, models, decisions, and proposed solutions so that all of the key stakeholders can stay apprised of and comment on the potential impact of proposed Agency actions on the national consumer financial market. Second, the Agencies should require financial institutions to share with the public as much information as possible regarding their AI systems and assessments of those systems to enable researchers and those impacted to evaluate the efficacy and impact of the systems.

- Engagement: The Agencies should stay engaged with a diverse group of key stakeholders, including civil rights organizations, consumer advocates, and impacted communities in order to receive ongoing input and feedback on these important decisions. The proposed solutions to AI risks are likely to have significant implications for borrowers and communities of color as well as other vulnerable communities, such as individuals with disabilities, families, and Limited English Proficiency borrowers. The Agencies should regularly engage with these communities and seek solutions that treat all borrowers and communities equitably.

DETAILED RESPONSES TO THE REQUEST FOR INFORMATION

Our organizations' recommendations focus on the Agencies' responsibility to ensure that AI risk analysis works hand in hand with fair lending risk analysis, and that efforts to address AI risks do not exclude or undermine efforts to promote fair lending and equitable outcomes. In particular, any new initiatives intended to address AI risks should be thoroughly reviewed for fair lending risks to avoid the potential for any illegal discriminatory treatment or effect for communities of color and other underserved communities. More specifically, our organizations provide the following analyses and recommendations in response to the request for information.

A. Explainability

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?*

---

[47] *See, e.g.,* Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing, *in* Conference on Fairness, Accountability, and Transparency 33, 39 (2020) (stressing the importance of "standpoint diversity," as algorithm development implicitly encodes developer assumptions of which they may not be aware). *See also* Model Risk Management Guidance at 4 (stating that "[a] guiding principle for managing model risk is 'effective challenge' of models, that is, critical analysis by objective, informed parties who can identify model limitations and assumptions and produce appropriate changes").

[48] *See, e.g.,* Steve Lohr, Facial Recognition is Accurate, if You're a White Guy, New York Times (Feb. 9, 2018) (explaining how Joy Buolamwini, a Black computer scientist, discovered that facial recognition worked well for her White friends but not for her).

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.*
3. *For which uses of AI is lack of explainability more of a challenge? Please describe those challenges in detail. How do financial institutions account for and manage the varied challenges and risks posed by different uses?*

Explainability Risks

Colloquial definitions of explainability and definitions in AI/ML often diverge, in part because when humans make decisions, their reasoning is based on causal heuristics, but in AI/ML, reasons for decisions are typically only associative/pattern matching. At a high level, explainability in the context of AI/ML models generally refers to the ability to understand how a model produced an outcome for an individual given the input variables.[49] Interpretability, on the other hand, often refers to the ability to understand how models operate at a population level. Both explainability and interpretability are necessary to have confidence that models are non-discriminatory and equitable; accordingly, the Agencies should ensure that institutions do not use models unless they have robust methods for understanding how those models work across multiple dimensions. This is critical as a recent study revealed that 65% of respondent companies were not able to explain, with specificity, how AI decisions predict certain outcomes. Moreover, 68% of companies in the study reported they have insufficient mechanisms in place to comply with existing regulations.[50]

If a model is not interpretable, it may be difficult or impossible to assess whether variables in the model are functioning as proxies for protected classes because their predictive value is attributable solely or largely to their correlation with a protected characteristic, which would be illegal and potentially constitute disparate treatment. Moreover, creditor practices that have a disproportionate negative impact on a prohibited basis must meet a legitimate business need.[51] If a model is not interpretable, it may not be possible to establish that variables in that model actually advance a legitimate business need (rather than, for example, functioning as close proxies for protected characteristics). Even if a model does not contain proxies and meets a legitimate business need, interpretability limitations may hinder the ability of institutions to conduct fair lending testing of such models. For example, while it is generally possible to assess whether such models cause disproportionate negative impacts, it can be more difficult to effectively identify whether less discriminatory versions of the model exist that meet the entity's legitimate need.

[49] Generally, this Comment Letter uses the term "explainable" to mean local explanations, and "interpretable" to mean global explanation. *See* Leilani H. Gilpin et al., Explaining Explanations: An Overview of Interpretability of Machine Learning, 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018.
[50] Solomon Radley, The State of Responsible AI: 2021 (May 2021).
[51] 12 C.F.R. Part 1002, Supp. I, ¶ 6(a)-2.

Moreover, explainability for individual decisions is important for generating adverse action reasons in accordance with ECOA and Regulation B.[52] Regulation B requires that creditors provide adverse action notices to credit applicants that disclose the principal reasons for denial or adverse action.[53] The disclosed reasons must relate to and accurately describe the factors the creditor considered.[54] This requirement is motivated by consumer protection concerns regarding transparency in credit decisionmaking and preventing unlawful discrimination. Adverse action notices that result from inexplicable models will not be helpful or actionable. Unfortunately, a recent CFPB blog post regarding the use of AI/ML models when providing adverse action notices seemed to emphasize the "flexibility" of the regulation rather than ensuring that AI providers and users adhere to the letter and spirit of ECOA, which was meant to ensure that consumers could understand the credit denials that impact their lives.[55] Our organizations believe that the complications raised by AI/ML models do not relieve creditors of their obligations to provide reasons that "relate to and accurately describe the factors actually considered or scored by a creditor."[56] Accordingly, the CFPB should make clear that creditors using AI/ML models must be able to generate adverse action notices that reliably produce consistent, specific reasons that consumers can understand and respond to, as appropriate. As the OCC has emphasized, addressing fair lending risks requires an effective explanation or explainability method, regardless of the model type used: "[b]ank management should be able to explain and defend underwriting and modeling decisions."[57]

AI/ML models and approaches vary significantly in terms of explainability, interpretability, and recourse (ability of individuals to respond to and improve their outcomes based on explanations).[58] Given creditors' legal obligations, entities should not use AI/ML models for which they cannot reasonably assess fair lending risks at a population level and for individual consumers.

Moreover, the current regulatory guidance should be updated to reflect the complexities of generating borrower-specific reasons for AI/ML models. The Official Commentary to Regulation B provides two example methods for identifying adverse action reasons based on credit scoring systems. One is to identify the factors for which an applicant's score fell furthest below the average score for each of those factors "achieved by applicants whose total score was at or slightly above the minimum passing score." Another method is to identify the same with respect

---

[52] Adverse action notices are also required under the Fair Credit Reporting Act ("FCRA") for certain events including adverse actions as defined in ECOA, and when adverse actions in employment, insurance, and certain other contexts are taken on the basis of information in a consumer report. *See* 15 U.S.C. 1681a(k).
[53] 12 C.F.R. Part 1002, Supp. I, ¶ 9(b).
[54] *Ibid.*
[55] Patrice Alexander Ficklin, Tom Pahl, and Paul Watkins, Innovation Spotlight: Providing Adverse Action Notices When Using AI/ML Models, CFPB Blog (July 7, 2020).
[56] 12 C.F.R. Part 1002, Supp. I, ¶ 9(b)(2)-2.
[57] OCC, Semiannual Risk Perspective, 23 (Spring 2019).
[58] While not the focus of this RFI, the Agencies should consider how best to move beyond purely providing explanations to also ensuring that the variables used and explanations provided allow consumers to act to improve their financial situation and future access, especially when denied financial services. *See* Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera, A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects, arXiv (Mar. 2021) (providing a survey of the literature).

to factors "achieved by all applicants."[59] Other methods that produce "results substantially similar to" either of those illustrative methods are also acceptable.[60]

It is not clear that these illustrative methods would actually identify valuable adverse action reasons in many contexts. In particular, there are likely circumstances where using "all applicants" as a baseline would not in fact reliably identify principal adverse action reasons. For example, an applicant's score may fall significantly below the average score for a factor achieved by all applicants, but depending on the model, that metric may not be informative regarding whether improvements to that factor—or how feasible it is for applicants to reasonably improve that factor—would increase the likelihood of approval. Other methods that produce results "substantially similar" to this method would also be acceptable under the Commentary, but could raise the same concerns. At the same time, it is not clear that identifying the factors for which an applicant's score fell furthest below "the average score for each of those factors achieved by" a reference group of applicants reliably identifies reasons generated by AI/ML models. Given how those models operate, identifying an average score for factors may not be a meaningful metric.

Regardless, this Commentary should be updated to reflect complexities introduced by AI/ML models and ensure that entities generate and provide clear notices to consumers.

Risk Management

The EU Proposed Regulation's approach to explainability may be a useful starting point to mitigating explainability risks. The preamble to the proposed regulation states the importance of explainability and transparency, including in relation to non-discrimination.[61] The proposed regulation would require that providers of high-risk AI systems provide users with clear instructions for use, including specifications for the input data, or any other relevant information in terms of the training, validation, and testing data sets used, taking into account the intended purpose of the AI system; and the required human oversight measures, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users.[62]

Recommendations

Accordingly, our organizations recommend the following:

- Staff Expertise: The Agencies should develop their own in-house expertise to test and understand explainability methods specific to the needs of consumer financial services.

- Policy Statement regarding Supervisory Expectations for Interpretability and Explainability: The Agencies should release a policy statement that clearly explains their expectations for the interpretability and explainability of AI/ML models as well as descriptions of best-in-class practices.

[59] 12 C.F.R. Part 1002, Supp. I, ¶ 9(b)(2)-5.
[60] *Ibid*.
[61] EU Proposed Regulation at Recital 47.
[62] *Id*. at Title III, Ch. 2, Art. 13.

- Policy Statement regarding Discrimination Risks: The Agencies should issue a policy statement clarifying that models that are not transparent and comprehensible raise serious discrimination risks, in addition to risks related to generating adverse action reasons.

- Updates to Regulation B and the Commentary: The CFPB should update Regulation B and the relevant Commentary to clarify that methods that do not reliably return sensible and accurate results do not comply with Regulation B. In addition, the CFPB should immediately clarify that the complications raised by AI/ML models do not relieve creditors of their obligations to provide adverse action reasons that relate to and accurately describe the factors actually considered or scored by a creditor.

- Supervisory and Enforcement Action: Consistent with the Uniform Interagency Consumer Compliance Rating System,[63] the Agencies should issue Matters Requiring Attention to financial institutions that use AI/ML if the Compliance Management System is inadequate in identifying and managing the risk, including using appropriate explainability techniques, model documentation, and fair lending documentation. If Compliance Management System weaknesses result in violations of law, including violations of the Fair Housing Act, ECOA, or Regulation B, then the Agencies should take appropriate enforcement actions, including referrals to the Department of Justice, non-public Memoranda of Understanding, or public enforcement actions.

B. Risks from Broader or More Intensive Data Processing and Usage

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

Data Risks

Data quality is critically important to the development of an AI model that is non-discriminatory, equitable, and effective for its intended purpose.[64] There are several challenges related to data quality in the context of fair lending and the effective use of AI in financial services.

Under-inclusive/Sample Bias: The data may be under-inclusive or may reflect sample bias. That is, the data may not fairly represent the intended populations. For example, in many instances, people of color are disproportionately missing from credit data, in part because they often live in credit deserts and disproportionately access financial services from non-traditional, alternative

[63] FFIEC, Uniform Interagency Consumer Compliance Rating System (Nov. 7, 2016).
[64] *See* Model Risk Management Guidance at 6 (stating that "[t]he data and other information used to develop a model are of critical importance; there should be rigorous assessment of data quality and relevance, and appropriate documentation").

credit providers such as payday lenders, check cashers, and title money lenders, which often do not report payments to consumer reporting agencies. Studies have shown that Black and Latino Americans are more likely than White or Asian Americans to be "credit invisible" or to have unscored records.[65]

Historical Bias: The data may reflect historical bias or inequality. For example, concerns have been raised about AI systems based on appraisal data, which may reflect historical biases due to the HOLC maps and other biases. A 2018 Brookings Institution study found that homes in majority Black neighborhoods were appraised for 23 percent less than properties in mostly White neighborhoods, even after controlling for home features and neighborhood amenities, which raises questions about the appropriateness of the data.[66] Moreover, even if the data excludes race and other protected characteristics, the risk may still be present through proxies or historical bias that is integrated into the model. If left unmitigated, this historical bias generally leads to feature bias in AI algorithms, as the algorithms would miss out on features that are truly predictive if the history behind the data already excluded such features.

Inappropriate Use of Race or Other Protected Characteristics: The data may fail to use race or other protected class data appropriately. First, the data may inappropriately and illegally include protected class or proxy data in the model.[67] With limited explicit exceptions, it is a violation of the fair lending laws' prohibitions against overt, intentional discrimination to use a protected class as a variable in a credit scoring or pricing model.[68] This is equally true for close proxies, such as zip code, geographic location, or language preference.[69] In addition, the data may inappropriately *exclude* the data needed to test the model's outcomes for discrimination risks. While race or other protected class data may not be appropriate to use in the model, this information is necessary for testing whether the model causes disproportionate adverse impacts on protected classes and for conducting an analysis of less discriminatory alternatives.

---

[65] *See, e.g*, Kenneth P. Brevoort, Phillipp Grimm, and Michelle Kambara, Data Point: Credit Invisibles, CFPB Office of Research, 6 (May 2015).

[66] Andre Perry, Jonathan Rothwell, and David Harshbarger, The Devaluation of Assets in Black Neighborhoods, The Brookings Institution Metropolitan Policy Program (Nov. 2018). *See also* Junia Howell and Elizabeth Korver-Glen, Neighborhoods, Race, and the Twenty-first Century Housing Appraisal Industry, 4 Sociology of Race and Ethnicity 473 (2018) (finding substantial differences in home values in communities of color even after controlling for home features, neighborhood amenities, socioeconomic status and consumer demand).

[67] Relatedly, the Agencies may want to consider the risks of chartered industrial loan companies ("ILCs") that will be able to make greater use of their broad insights into the lives of consumers. Access to purchase history and in-platform search queries may allow ILCs to more easily use data points that are proxies for discriminatory data points. *See* Congressional Research Service, CRS Report for Congress: Financial Privacy Laws Affecting Sharing of Customer Information Among Affiliated Institutions (Feb. 23, 2005).

[68] *See* Regulation B, 12 C.F.R. Part 1002, Supp. I, ¶ 2(p)–4 ("Besides age, no other prohibited basis may be used as a variable."); FFIEC, Interagency Fair Lending Examination Procedures at 8 (Aug. 2009) (explaining that overt discrimination includes using "variables in a credit scoring system that constitute a basis or factor prohibited by Regulation B or, for residential loan scoring systems, the FHAct"); OCC, Bulletin 97-24, Appendix, "Safety and Soundness and Compliance Issues on Credit Scoring Models" (1997) (noting that "a creditor cannot use a credit scoring system that assigns various points based on the applicant's race, national origin, or any other prohibited basis," with an exception for age).

[69] *See* OCC, Bulletin 97-24 at 10 ("Moreover, factors linked so closely to a prohibited basis that they may actually serve as proxies for that basis cannot be used to segment the population.").

Alternative Data: Traditional credit history scores reflect immense racial disparities due to extensive historical and ongoing discrimination.[70] Black and Latino consumers are less likely to have credit scores in the first place, limiting their access to financial services.[71] There is an obvious need for better, fairer, and more inclusive measures of creditworthiness.[72]

New data sources can help. But caution is in order: Not all kinds of data will lead to more equitable outcomes, and some can even introduce their own new harms.[73] Fringe alternative data such as online searches, social media history, and colleges attended can easily become proxies for protected characteristics, may be prone to inaccuracies that are difficult or impossible for impacted people to fix, and may reflect long-standing inequities. On the other hand, recent research indicates that more traditional alternative data, such as cash flow data, hold promise for helping borrowers who might otherwise face constraints on their ability to access credit.[74] Moreover, a recent Interagency Statement observed that "[c]ash flow data are specific to the borrower and generally derived from reliable sources, such as bank account records, which may help ensure the data's accuracy. Consumers can expressly permit access to their cash flow data, which enhances transparency and consumers' control over the data."[75]

Risk Management

To manage data risks, the EU Proposed Regulation may be a useful starting point. The proposed regulation states the importance of robust data governance with respect to fair lending: "High data quality is essential for the performance of many AI systems, especially when techniques involving the training of models are used, with a view to ensure that the high-risk AI system performs as intended and safely and it does not become the source of **discrimination** prohibited by [European] Union law."[76] Given the critically-important role that data plays in developing non-discriminatory and equitable AI systems, the proposed regulation would impose several requirements, including the requirement that data sets be examined in view of possible bias. Likewise, New York's Digital Fairness Act would require government agencies seeking to procure or utilize AI/ML systems to publish assessments that contain detailed descriptions of training data prior to undergoing a public comment and agency approval process and create AI/ML use policies that state the type of data used by the system and how it is "generated, collected, and processed."[77]

---

[70] *See* National Consumer Law Center, Past Imperfect: How Credit Scores and Other Analytics "Bake In" and Perpetuate Past Discrimination (May 2016); Jung Hyun Choi, Alanna McCargo, Michael Neal, Laurie Goodman, and Caitlin Young, Explaining the Black-White Homeownership Gap: A Closer Look at Disparities across Local Markets, Urban Institute (Oct. 10, 2019).
[71] *See* Kenneth P. Brevoort, Philipp Grimm, and Michelle Kambara, Data Point: Credit Invisibles, CFPB (May 2015).
[72] *See* Chi Chi Wu, Reparations, Race, and Reputation in Credit: Rethinking the Relationship between Credit Scores and Reports with Black Communities (Aug. 7, 2020).
[73] *See* Testimony of Aaron Reike, Managing Director, Upturn, Hearing: Examining the Use of Alternative Data in Underwriting and Credit Scoring to Expand Access to Credit, Task Force on Financial Technology, U.S. House Committee on Financial Services (July 25, 2019).
[74] *See* FinRegLab, The Use of Cash-Flow Data in Underwriting Credit (July 2019).
[75] *See* Federal Reserve, CFPB, FDIC, OCC, NCUA, Interagency Statement on the Use of Alternative Data in Credit Underwriting (Dec. 3, 2019).
[76] EU Proposed Regulation at Recital 44.
[77] AB-A06042, Gen. Assemb., Reg. Sess. (N.Y. 2021-2022) ("NY State Digital Fairness Act").

In addition to the EU Proposed Regulation and the New York Digital Fairness Act, the Agencies should review recent scholarship, which has proposed the use of a standardized process for documenting and archiving datasets.[78] For example, in the electronics industry, every component is accompanied by a datasheet that describes its operating characteristics, test results, recommended uses, and other information. Similarly, datasheets for datasets would facilitate better communication between dataset creators and dataset consumers, and encourage the prioritization of transparency and accountability.

Recommendations

Accordingly, our organizations recommend the following:

- Policy Statement regarding Data Governance and Fair Lending: The Agencies should issue a joint policy statement explaining the importance of quality, inclusive, and unbiased data to developing AI solutions that ensure compliance with fair lending laws and that promote equitable principles. In addition, the policy statement should explain the Agencies' supervisory expectations with respect to data governance for financial institutions that use AI and the institutions' obligations with respect to comprehensive reproducibility, e.g., documentation and archiving of data used in training.

- Public Research: The Agencies should encourage public research on data risks and the evaluation of potential risk mitigation techniques.

C. Overfitting

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 Risks

A typical indicator of poor data quality is when "overfitting" occurs. In general, overfitting is when a model performs well on the training data but does a poor job on new data that was not used in training. Financial institutions use the conventional definition of overfitting that emphasizes the performance or accuracy of the machine learning model on an underlying "true" data distribution, which is usually estimated by measuring the model's performance on a hold-out test or validation set. Overfitting results when a machine learning algorithm fails to reproduce its training performance on a test or validation dataset, or when deployed in a production environment. This risk of overfitting is a challenge for AI adoption, as the quality of

[78] *See* Timnit Gebru and Jamie Morgenstern et al., Datasheets for Datasets (Mar. 19, 2020). *See also* Sasu Makinen, Henrik Skogstrom, Eero Laaksonen, and Tommi Mikkonen, Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help (Mar. 16, 2021); Pukrit Agrawal et al., Data Platform for Machine Learning (2019); Margaret Mitchell and Simone Wu et al., Model Cards for Model Reporting (Jan. 2019); Rachel Tatman, Jake VanderPlas, and Sohier Dane, A Practical Taxonomy of Reproducibility for Machine Learning Research (2018).

data within financial institutions is not always representative of real data due to inconsistent data definitions, data silos across jurisdictions, and the use of multiple data systems.[79]

Risk Management

There are many approaches to mitigating the risk of overfitting, and the Agencies should carefully review the pros and cons of each approach before providing the industry with guidance on best practices. Some common risk mitigation techniques include:

- Bias-variance Tradeoff: Bias-variance tradeoff is a classical theory often used by machine learning practitioners as a guide for identifying and mitigating overfitting.[80] Bias (used here to refer to model bias generally, not protected-class bias) is the error due to an algorithm finding an imperfect model and it is low when the algorithm learns close to the ground truth. Variance is the error due to lack of data, and a complex model with more parameters to tune will require more data. Hence there is a bias-variance tradeoff with a fixed amount of data: a complex model with many parameters could be accurate, but there is limited amount of quality data to estimate it appropriately (low bias and high variance scenario); or a simple model with low accuracy can estimate its parameters well given the data (high bias and low variance scenario). This tradeoff could be further complicated by noise, the error due to exclusion of relevant features from the model, or because the process generating the underlying data is inherently stochastic. However, there are examples of unusual phenomena that challenge this classical tradeoff approach,[81] and therefore there should be robust mechanisms for mitigating overfitting risks.[82]

- Balanced Data Design: Block or stratified data sampling strategies can be used to ensure that both the training data and the test data have similar patterns.

- Validation Set: In addition to using test-train split to estimate model performance, a validation set—either appropriately randomly sampled or time-series based—accessible only to the model risk management teams of the financial institutions can be used to validate model performance. This can allow independent measurement of the effectiveness of overfitting prevention strategies.

- Documentation of Overfitting Prevention Strategies: Financial institutions can document the approach they use to mitigate overfitting so that an independent reviewer can make an informed decision about the appropriateness of the approach.

- Model Drift Monitoring: Financial institutions can have an end-to-end model deployment process that includes monitoring drifts in model predictions and model features post-deployment. This may trigger model retraining if the drifts suggest that data in the

[79] Bart van Liebergen, Machine Learning: A Revolution in Risk Management and Compliance?, The CAPCO Institute Journal of Financial Transformation (2017).
[80] Mikhail Belkin et al., Reconciling Modern Machine Learning Practice and the Bias-variance Tradeoff, arXiv (Sept. 2019).
[81] Chenlei Fang et al., The Overfitting Iceberg, Carnegie Mellon University ML Blog (Aug. 31, 2020).
[82] Xue Ying, An Overview of Overfitting and Its Solutions, Journal of Physics: Conference Series 1168 (2018).

production environment is coming from a population that may be different from that which generated the training and test data.

Recommendations

Accordingly, our organizations recommend the following:

- Policy Statement regarding Overfitting: The Agencies should evaluate the available risk mitigation techniques and then issue a joint policy statement explaining the risks of overfitting, the Agencies' supervisory expectations, and best practices for risk mitigation techniques.

- Public Research: The Agencies should encourage public research on the risks of overfitting and the evaluation of potential risk mitigation techniques.

D. Dynamic Updating

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?

Dynamic Updating

Monitoring AI solutions allows financial institutions to detect changes between patterns of the data used to develop the deployed model (i.e., development environment) and patterns of the customer data used to make business decisions post-deployment (i.e., production environment). The dynamics of the business environment can then inform an automated or a semi-automated decision to dynamically update the AI models depending on how material the change is. It should be noted, however, that dynamic updating is not the norm for most underwriting models. For example, the credit scoring models and Automated Underwriting Systems used by the GSEs do not use dynamic updating.

Risk Management

Dynamic updating comes with challenges and there are many approaches to mitigating these challenges. The Agencies should carefully review the pros and cons of each approach before providing the industry with guidance on best practices. We note some of the common frameworks for addressing the challenges of dynamic updating here.

- Model Risk Management: Model Risk Management is an established framework that financial institutions use to manage risks related to the operation of models. This framework consists of comparing model predictions on training sets with predictions post-deployment and using summary statistics to make statistical decisions on the level of the drift in the underlying datasets, i.e., development datasets (test and training sets) and

production datasets.[83] Population stability index ("PSI") and characteristic stability index ("CSI") are common summary statistics that financial institutions use for model monitoring. While PSI measures drifts in predictions such as credit score and default risks, CSI measures drifts in features used in the underlying model or algorithm. Both PSI and CSI are estimation metrics designed to overcome some of the challenges of dynamic updating.

- Machine Learning Observability: Machine Learning Observability ("MLO") is an area of AI that is broader in scope than model risk monitoring and is generally interested in what could be inferred from the model's predictions, production feature and development feature data drifts, and whether the same inputs over time produce different outputs when fed into the underlying algorithms. Findings from MLO are then used to build workflows that are used to dynamically update the machine learning models.

Key challenges to dynamic updating of AI models (using frameworks like MLO) include but are not limited to:

- Delayed Observation or Lack of Ground Truth: The actual outcome of model prediction often takes a while to observe in many financial applications of AI. For example, loan terms could vary from one month (e.g., payday loans) to five years (e.g., auto loans) or even up to thirty years (e.g., mortgage loans) and as such real time or online updating of credit scoring solutions is not feasible. The actual ground truth may never be observed in extreme situations.

- Sequential Decision Making and the Dynamic Nature of Outcomes: As with the delayed observation challenge, there also remains an issue related to the non-static nature of decisions in financial institutions. For instance, credit is perceived as a static prediction, but in reality, people will engage with credit repeatedly (e.g., apply for multiple credit cards), and prior engagements with lenders impact future predictions. Focusing on tracking single point-in-time concepts of fairness or model effectiveness may not help regulators understand the long-term impacts of AI-systems on impacted populations.[84]

- Missing Data Problem: There are sophisticated statistical methods to handle missing data depending on the pattern of missingness. Missingness patterns in production data often diverge from those found in development data (i.e., training, test, and validation data), and applying the same missing data models to both production and development data may significantly render stability indices such as CSI and PSI meaningless.

- Lack of Rigorous Thresholds: The industry does not have a rigorous standard for what constitutes model or feature drifts. For example, the rule of thumb for deciding how far a model outcome or a model feature has drifted is: when PSI is below 0.1, then the drift is slight; when PSI is between 0.1 and 0.2, then the drift is minor; and when PSI is above

[83] Bilal Yurdakul and Joshua Naranjo, Statistical Properties of the Population Stability Index, Journal of Risk Model Validation, Vol. 14(4) (Aug. 14, 2019).
[84] Lydia T. Liu et al., Delayed Impact of Fair Machine Learning, Proceedings of the 35th International Conference on Machine Learning (2018).

0.2, then the drift is significant.[85] While it is not clear if these rules of thumbs may be applied to drifts in model features, many applications of these thresholds in AI solutions show that they are not reliable measures of drifts.[86]

- Information Loss: A typical AI algorithm usually has hundreds of features, and computing CSI for each feature can result in numerous quantitative results whereas a single summary statistic like PSI is often desirable. However, many credit scoring applications of AI result in continuous outputs, and using PSI on continuous outputs requires variable categorization that may lead to loss of information.

In addition to the specific techniques mentioned above, the EU Proposed Regulation's approach to post-deployment monitoring more broadly may be useful to consider. The preamble to the proposed regulation clearly states the importance of ongoing compliance for high-risk AI systems, including those that evaluate creditworthiness. The preamble states that a sound post-deployment monitoring plan is "key to ensur[ing] that the possible risks emerging from AI systems which continue to 'learn' after being put into production can be more efficiently and timely addressed."[87] More specifically, the proposed regulation would require providers of high-risk AI systems to implement a post-deployment monitoring plan that would (i) actively and systematically collect, document, and analyze relevant data provided by users or collected through other sources on the performance of high-risk AI systems throughout their lifetime; and (ii) allow the provider to evaluate the continuous compliance of AI systems.[88]

Recommendations

Accordingly, our organizations recommend the following:

- Policy Statement regarding Dynamic Updating: The Agencies should carefully evaluate the available risk mitigation techniques and then issue a joint policy statement explaining the risks of dynamic updating, the Agencies' supervisory expectations, and best practices for risk mitigation techniques.

- Public Research: The Agencies should encourage public research on the risks of dynamic updating and the evaluation of potential risk mitigation techniques. For example, while vintage analysis and roll rate analyses may share some of the challenges of dynamic updating listed above, researchers should investigate how they can be scaled to AI solutions in financial applications.

[85] Alec Zhixiao Lin, Examining Distributional Shifts by Using Population Stability Index (PSI) for Model Validation and Diagnosis (2017).
[86] Ross Taplin and Clive Hunt, The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring, Risks, MDPI (May 6, 2019).
[87] EU Proposed Regulation at Recital 78.
[88] *Id*. at Title VIII, Ch. 1, Art. 61.

E. Fair Lending

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.

As noted above, the traditional disparate treatment and disparate impact frameworks translate well to assessments of consumer financial services lending models, including AI/ML models. The existing systems for testing models for discrimination include: (1) ensuring that models do not include protected classes or close proxies for protected classes, for example as variables or segmentations; and (2) assessing whether facially-neutral models are likely to disproportionately lead to negative outcomes for a protected class, and, if such negative impacts exist, ensuring the models serve legitimate business needs and evaluating whether changes to the models would result in less of a disparate impact while maintaining model performance.

In addition, there are several algorithmic approaches to improving model fairness, particularly for AI models. The Agencies should carefully review the pros and cons of each of these methods before providing the industry with guidance on best practices. We note some of the common techniques here and recommend that the agencies carefully evaluate each of them.

- Pre-Processing: One of the core technical problems in AI is sometimes referred to as "Garbage in, garbage out": modern AI algorithms are "trained" on the basis of what's happened in the past. If the data on which an AI is trained is not complete, balanced, representative, or selected appropriately, it can be a major source of bias. Moreover, the data itself may hold inherent biases that reflect discriminatory practices in the society. There are proposed tools[89] and techniques[90] for attempting to quantify the bias in a data set and either reduce or nullify it. The Agencies should explore these methods, support research into credit-finance specific data de-biasing, and provide guidance on whether they should be used to supplement human reviews of variables for potential discriminatory effect.

- In-Processing: In-processing techniques direct a learning model to optimize for prediction accuracy but within fairness constraints imposed by the model developer. Some studies suggest that in-processing techniques can mitigate model bias while maintaining good overall accuracy.[91] Other studies have found that in-processing mitigation techniques can satisfy static fairness criteria but can hurt long-term outcomes.[92] In addition, in-processing techniques often require replacing a deployed model with a new algorithm, which can take considerable effort to modify and maintain.[93] Here too, the Agencies should explore these methods, support credit-finance specific research on in-processing

[89] IBM Cloud Pak for Data, Debiasing Options.
[90] Ulrich AIvodji et al., Adversarial Training Approach for Local Data Debiasing, arXiv (Sept. 3, 2020).
[91] Cristina Gorrostieta et al., Gender De-biasing in Speech Emotion Recognition, ISCA Interspeech (2019).
[92] Lydia T. Liu et al., Delayed Impact of Fair Machine Learning (2018).
[93] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann, Fairness in Credit Scoring: Assessment, Implementation and Profit Implications, arXiv (2021).

fairness techniques, and provide guidance and clarity on whether and how they could be effectively used.

- Post-Processing: Post-processing techniques apply adjustments to models after the model is built to meet some specific fairness constraint.[94] For instance, reject option classification is a technique that labels instances belonging to disfavored and favored groups after the algorithm is built to optimize some fairness criteria at the decision boundary (e.g., the score boundary between receiving a loan or not).[95] Post-processing techniques offer the advantage that they are among the easiest to implement, requiring no knowledge or access to the internal workings of the model, but sometimes improve fairness at a higher cost to accuracy than other techniques.[96] In addition, post-processing often requires use of protected class data during production, which may raise concerns under ECOA.

Pre-processing and in-processing fairness enhancement techniques use protected class data during model training, but none of them uses that information while actually scoring applications in production. This raises the question of what ways the awareness or use of protected class data during training is legal. If protected class data is being used for a salutary purpose during model training—such as to do a better job of approving loan applicants from historically disadvantaged groups—there would seem to be a strong policy rationale for permitting it, but there is no regulatory guidance on this subject.

Considerations That Attend the Search for Less Discriminatory Alternatives: The fairness enhancing techniques above are not exhaustive; there are more approaches to mitigating disparate impact that are often effective. One approach that is commonly used in traditional modeling, and that remains viable with machine learning models, is simply to change the model's threshold required for acceptance. This may favorably change the demographic balance between protected classes at little to no cost to the business. In particular, if there are relatively more borrowers of color, for example, just beyond the initial cutoff proposed, it may be possible to decrease disparate impact by changing the cutoff slightly without substantially affecting the business's expected profitability.

Alternatively, the structure of the AI model can be changed, either by choosing a different kind of model or by changing the architecture of the model. For example, the underlying algorithm could be changed from an XGBoost algorithm to a random forest algorithm (which are two types of machine learning algorithms commonly used in credit scoring). Or, if the same model type is used, the model's parameters can be changed, including those that control the internal complexity of the model. Another method, "feature selection," where certain variables that are found to be disproportionate drivers of disparate impact are dropped from the model and other variables that cause less disparity are added, can be effective at removing the impact of especially problematic variables. An advantage of this methodology is its widespread

[94] Moritz Hardt, Eric Price, and Nathan Srebro, Equality of Opportunity in Supervised Learning, Advances in Neural Information Processing Systems (2016).
[95] Faisal Kamiran, Asim Karim, and Xiangliang Zhang, Decision Theory for Discrimination-aware Classification, IEEE 12th International Conference on Data Mining (2012).
[96] Nikita Kozodoi, Johannes Jacob, and Stefan Lessmann, Fairness in Credit Scoring: Assessment, Implementation and Profit Implications, arXiv (Mar. 4, 2021).

acceptance, in that it has been used by lenders for decades to address disparate impact in traditional models. Further, an additional benefit of both feature selection and parameter tuning is that they do not require that the model itself have access to any protected class information at any point during training. Instead, the disparate impact is evaluated only after the model has been trained. Importantly, none of the methods discussed in this section is mutually exclusive; feature selection, parameter optimization, and changing score cutoffs can be used along with, or in place of, a combination of pre-processing, in-processing, and post-processing techniques.

An important consideration raised by the use of these fairness enhancing techniques is that these algorithmic methods must be deliberately applied by humans. After applying these techniques, a human must then choose among the resulting AI models, to balance between fairness and business objectives. These choices are manageable, and many institutions currently have internal guidelines on how they should be made. But public guidance would be useful to ensure that institutions are properly taking into account their anti-discrimination obligations.

Recommendations

Accordingly, our organizations recommend the following:

- Policy Statement regarding AI-based Fair Lending Compliance: The Agencies should carefully assess the techniques used to evaluate AI-based credit determinations compliance with fair lending laws and then issue a joint policy statement explaining the risks, the Agencies' supervisory expectations, and best practices. The Agencies, or the CFPB independently, should issue a white paper describing the techniques that they use to assess discrimination risks posed by institutions' models.

- Public Research: The Agencies should encourage public research on the evaluation of techniques used to evaluate the fair lending risks associated with AI-based credit determinations.

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?

There is a High Risk That AI Will Amplify Historical Discrimination

As explained above, right now, the United States is at a crossroads in determining whether to develop equitable AI systems that serve and uplift the whole of the national financial services market, or systems that perpetuate, amplify, and even accelerate old discriminatory patterns. For much of America's history, communities of color were systematically excluded from economic opportunities through explicit government policy decisions that inculcated an inappropriate and unfounded association between race and risk into the nation's housing and financial markets.[97]

[97] *See* Lisa Rice, "The Fair Housing Act: A Tool for Expanding Access to Quality Credit," The Fight for Fair Housing: Causes, Consequences, and Future Implications of the 1968 Federal Fair Housing Act (Gregory Squires,

Even today, there is a separate and unequal financial services landscape in which mainstream creditors are concentrated in predominantly White communities and non-traditional, higher-cost lenders, such as payday lenders, check cashers, and title money lenders, are hyper-concentrated in predominantly Black and Latino communities.[98] These discriminatory practices and several others have created distinct advantages for White families, leading to massive wealth, homeownership, and credit gaps between White families and families of color.[99]

While AI/ML systems may be relatively new innovations, the building blocks for these models are tainted with historical bias and carry a high risk of discriminatory and inequitable outcomes. These systems can have a disparate impact on people and communities of color, particularly with respect to credit, because they reflect the dual credit market that resulted from our country's long history of discrimination. For example, credit scoring systems have been found to discriminate against people of color.[100] Risk-based pricing systems can perpetuate bias as well. In a Berkeley study, researchers found that certain algorithmic-based pricing systems discriminated against Black and Latino people, overcharging them by more than $765 million per year.[101] Finally, entities like Facebook that play an important gatekeeping role in the housing and credit markets can offer marketing and advertising services based on models, some of which have also been the focus of civil rights lawsuits.[102] In short, much like the historical HOLC maps, AI systems can penalize people simply because of the communities in which they live and associate and the types of risky credit historically targeted to those communities.

There Are Specific Risks Related to AI Data and Models

Data Risks

Building non-discriminatory and equitable AI systems requires the use of quality, reliable, robust data that truly reflects the patterns and behaviors of the people the models are designed to assess. While AI systems can be powerful and innovative, they can only see the patterns that exist in the data. Oftentimes, that data is reflective of the structural biases that pervade our society. There are several risks related to the use of data in AI systems, including that: (i) data can be under-inclusive; (ii) data can reflect historical discrimination; (iii) data can fail to use race and other protected class information appropriately, and (iv) some alternative data may lead to

---

1st ed. 2017) (providing a detailed explanation of how federal race-based housing and credit policies promoted inequality). *See also* K. Steven Brown et al., Confronting Structural Racism in Research and Policy Analysis, Urban Institute (Feb. 2019); Richard Rothstein, The Color of Law: A Forgotten History of How Our Government Segregated America (2017).

[98] *See* Cheryl Young and Felipe Chacón, 50 Years After the Fair Housing Act – Inequality Lingers, Trulia (Apr. 19, 2018).

[99] *See* Neil Bhutta et al., Disparities in Wealth by Race and Ethnicity in the 2019 Survey of Consumer Finances, FEDS Notes, Board of Governors of the Federal Reserve System (Sept. 28, 2020); Heather Long and Andrew Van Dam, The Black-White Economic Divide Is as Wide as It Was in 1968, Washington Post (June 4, 2020); Bruce Mitchell and Juan Franco. HOLC "Redlining" Maps: The Persistent Structure Of Segregation And Economic Inequality, National Community Reinvestment Coalition (Feb. 2018).

[100] Sarah Ludwig, Credit Scores in America Perpetuate Racial Injustice. Here's Why, The Guardian (Oct. 13, 2015).

[101] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace, Consumer Lending Discrimination in the FinTech Era, UC Berkeley Public Law and Legal Theory Research Paper Series (Sept. 11, 2019).

[102] *See* National Fair Housing Alliance, Facebook Settlement (Mar. 19, 2019).

discriminatory or inequitable outcomes. Please see the response to Question 4 for more detail on data risks.

Model Risks

AI models can increase risk due to the models' greater complexity and their potential to exacerbate historical disparities and flaws in underlying data. There are several risks related to AI models, including: (i) models that are fundamentally flawed and discriminatory; (ii) models that result in discriminatory feedback loops; (iii) models that have not been tested for discriminatory outcomes; (iv) models that are not explainable or understandable; (v) models that contain the risk of overfitting; and (vi) risks related to dynamic models.

The Model Can Be Fundamentally Flawed and Discriminatory: AI systems can be designed in ways that are fundamentally flawed and result in discriminatory or inequitable outcomes.[103] For example, systems that allow users to exclude certain racial or ethnic groups can cause discrimination against protected groups and even enhance the different ways in which users can discriminate against people. The National Fair Housing Alliance, several of its member organizations, the ACLU, and other civil rights groups filed legal challenges against Facebook because the company allowed entities placing ads for housing, employment, and credit on Facebook's platform to target audiences based on protected class characteristics like race, national origin, and gender.[104] As a result of these legal challenges, Facebook had to make several structural changes to its advertising platform and AI systems.

AI systems that use a scoring system to determine ad placement can also generate discriminatory or inequitable outcomes. For example, a Harvard researcher found that Google searches for people with Black-identifying names turned up more ads suggestive of arrest records and/or criminal backgrounds than did ad searches using White-identifying names.[105] Researchers recommended that Google change the quality score of ads to discount for unwanted discriminatory or inequitable outcomes.

The Use of the Model Can Result in Discriminatory Feedback Loops: If not carefully designed, AI systems can inappropriately exacerbate discriminatory patterns. For example, if an ad features the image of a man, an AI system registering the content of the ad might skew the ad's delivery to men. Thus, more men are likely to see the ad. As more men click on the ad, the AI system might mis-perceive that men are more likely to be interested in seeing the ad than women and continue to over-skew the ad's delivery to even more men. If that ad pertains to offers of credit, this may result in "digital redlining," where women are not provided with the opportunity to learn about credit offers.[106] Similarly, predatory lending algorithms could result in digital reverse

[103] *See* Model Risk Management Guidance at 2 (stating that model risk occurs primarily for two reasons, including fundamental errors that can occur at any point from design through implementation).
[104] *See* National Fair Housing Alliance, Facebook Settlement (Mar. 19, 2019).
[105] LaTanya Sweeney, Discrimination in Online Ad Delivery, 11(3) ACMQueue (Apr. 2, 2013).
[106] *See* Carol A. Evans and Westra Miller, From Catalogs to Clicks: The Fair Lending Implications of Target, Internet Marketing, Federal Reserve Consumer Compliance Outlook (2019) (raising concerns about digital redlining that might render some advertisements invisible to certain users, disproportionately impacting users based on protected characteristics, such as race and sex).

redlining. For example, if an algorithm targets Black and Latino borrowers for predatory loans and they click on the ad, these borrowers will increasingly receive more of these ads.[107]

Similarly, the Berkeley study of risk-based pricing systems posited that algorithmic mortgage pricing may be overcharging Black and Latino borrowers based on reduced shopping activity.[108] However, reduced levels of mortgage loan shopping among Black and Latino borrowers may be caused by these borrowers disproportionately living in credit deserts, rather than caused by a greater risk of default. In this way, the structural inequities linked to residential segregation and the dual credit market serve as a discriminatory feedback loop that results in borrowers of color being charged more for credit when they pose no greater level of risk.

There Can be Failures in Adequately Testing Models for Discriminatory Outcomes: AI systems can be deployed without adequately testing them for discriminatory outcomes, which can result in consumer harm, violations of laws, and amplification of historically discriminatory lending patterns.[109] As noted, both ECOA and the Fair Housing Act prohibit disparate impact.[110] Consistent with existing law and policy, AI systems should be tested to determine whether facially-neutral models are likely to disproportionately lead to negative outcomes for a protected class. If such negative impacts exist, the AI systems should be reviewed to ensure that the models serve legitimate business needs and to determine whether any changes to the models would result in less of a disparate impact while maintaining model performance.[111]

Please see the following responses for more detailed information about other model risks:

- Explainability at Questions 1-3;
- Overfitting at Question 6; and
- Dynamic updating at Question 8.

Certain Policies and Techniques Can Mitigate AI Risk

There are significant risks that AI systems can result in discriminatory or inequitable outcomes, but the risks are not insurmountable. Following are recommendations as to how the Agencies, financial institutions, and tech companies can mitigate these risks.

---

[107] *See, e.g.,* Alexander D'Amour, et. al, Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies, Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency(2020).
[108] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace, Consumer-Lending Discrimination in the FinTech Era, UC Berkeley (2019).
[109] *See* Model Risk Management Guidance at 10 (stating that "[e]ffective model validation helps reduce model risk by identifying model errors, corrective actions, and appropriate use").
[110] *See, e.g.*, Regulation B, 12 C.F.R. Part 1002, Supp. I, ¶ 6(a)-2 (ECOA articulation); 24 C.F.R. § 100.500(c)(1) (Fair Housing Act articulation); 42 U.S.C. § 2000e-2(k) (Title VII articulation).
[111] *See* Relman Colfax PLLC, Fair Lending Monitorship of Upstart Network's Lending Model, Initial Report of the Independent Monitor, 7 (Apr. 14, 2021); Nicholas Schmidt and Bryce Stephens, An Introduction to Artificial Intelligence and Solutions to the Problems of Algorithmic Discrimination, 73(2) Quarterly Report 130, 141 (2019).

Agency Action

The Agencies can mitigate AI risk by using all of the tools in their toolbelt, including supervision, enforcement, policy, staffing, and support for public research.

- Non-discrimination and Equity: The Agencies should take the steps needed to ensure non-discriminatory and equitable outcomes for all who participate in the financial services market. Most importantly, the Agencies should define "model risk" to include the risk of discriminatory or inequitable outcomes for consumers, rather than just the risk of financial loss to a financial institution.[112] That is, the analysis of fair lending risk and equity should be integrated into all AI discussions and not treated as an afterthought. Throughout this document, we have demonstrated how the evaluation of fair lending risk and equity is an integral part of evaluating every aspect of AI risk.

- Action Plan: After review of the RFI responses, the Agencies should immediately issue a detailed Action Plan, which may include plans for a white paper regarding a proposed framework for the regulation of AI in financial services, a policy statement reminding financial institutions of their responsibilities under fair lending and other consumer protection laws, a policy statement describing the Agencies' expectations for financial institutions with respect to AI, a policy statement regarding the Agencies' methodologies for evaluating AI systems, a proposed regulation (including under the CFPB's UDAAP authority), and/or examination procedures.

- Robust Supervision and Enforcement/Accountability: The Agencies should conduct in-depth reviews of financial institutions' use of AI, including assessing compliance with fair lending laws.
    - Consistent with the Uniform Interagency Consumer Compliance Rating System[113] and the Model Risk Management Guidance,[114] the Agencies should ensure that financial institutions have appropriate Compliance Management Systems that effectively identify and control risks related to AI systems, including the risk of discriminatory or inequitable outcomes for consumers. The Compliance Management System should comprehensively cover the roles of board and senior management, policies and procedures, training, monitoring, and consumer complaint resolution. The extent and sophistication of the financial institution's Compliance Management System should align with the extent, sophistication, and risk associated with the financial institution's usage of the AI system, including the risk that the AI system could amplify historical patterns of discrimination in financial services.

---

[112] *See* Model Risk Management Guidance at 3 (defining "model risk" to focus on the financial institution rather than the consumer by stating that "[m]odel risk can lead to financial loss, poor business and strategic decision making, or damage to a bank's reputation").

[113] FFIEC, Uniform Interagency Consumer Compliance Rating System (Nov. 7, 2016) (stating that for purposes of a financial institution's consumer compliance rating, examiners will assess the financial institution's Compliance Management System based on the board and management oversight as well as the compliance program, which includes policies and procedures, training, monitoring, and complaint resolution). *See also* CFPB Bulletin 2020-01, Responsible Business Conduct: Self-Assessing, Self-Reporting, Remediating, and Cooperating (Mar. 6, 2020).

[114] Model Risk Management Guidance at 15 (stating that "[d]eveloping and maintaining strong governance, policies, and controls over the model risk management framework is fundamentally important to its effectiveness").

- Where a financial institution's use of AI indicates weaknesses in their Compliance Management System or violations of law, the Agencies should use all of the tools in their toolbelt to quickly address and prevent consumer harm, including issuing Matters Requiring Attention; entering into a non-public enforcement action, such as a Memorandum of Understanding; referring a pattern or practice of discrimination to the U.S. Department of Justice; or entering into a public enforcement action. The Agencies have already provided clear guidance (e.g., the Uniform Consumer Compliance Rating System) that financial institutions must appropriately identify, monitor, and address compliance risks, and the Agencies should not hesitate to act within the scope of their authority.
- Moreover, any new policies or initiatives related to AI should clearly state that the Agencies will hold financial institutions accountable for Compliance Management System weaknesses or violations of law.
- When possible, the Agencies should explain to the public the risks that they have observed and the actions taken in order to bolster the public's trust in robust oversight and provide clear examples to guide the industry.

- Actionable Policies: Existing civil rights laws and policies provide a framework for the Agencies to analyze fair lending risk in AI and to engage in supervisory or enforcement actions, where appropriate. That said, the Agencies can be more effective in ensuring consistent and effective compliance by setting clear and robust regulatory expectations regarding testing and ensuring models are non-discriminatory and equitable. The Agencies have been in learning mode for some time, which may have put the U.S. behind in advancing non-discriminatory and equitable technology in financial services. To retain our competitive edge in global society, the U.S. federal financial regulators should move quickly to issue actionable policy statements that clearly state their commitment to consumer protection and civil rights laws, including fair lending laws; insight into their supervisory expectations and methods; and useful guardrails and best practices. The time to act is now as the use of AI proliferates in every aspect of consumer financial services and has the potential for far-reaching adverse impacts for consumers of color and other protected groups. More specifically, the Agencies can be more effective in ensuring robust and consistent compliance by moving quickly to issue a clear policy statement on AI that:

  1. Defines "model risk" to include the risk of discriminatory or inequitable outcomes;
  2. Describes the risks that financial institutions should be aware of and control for;
  3. Sets clear standards for a financial institution's fair lending risk assessments, including:
     a. Discrimination testing and evaluation throughout the AI/ML model's conception, design, implementation, and use; and
     b. Information that must be detailed in the documentation of the financial institution's fair lending risk assessment, including:
        (i) What testing has been conducted and less discriminatory alternatives have been considered;

(ii) In-depth information regarding the data that was used to train the model, measures taken to ensure the data was representative and accurate, and the attributes used in the model and its target outcomes; and
(iii) Documentation on adverse action notices detailing the mechanism by which the adverse action notices are created and showing that the mechanism provides adverse action notices that reliably produce consistent and specific reasons that consumers can understand and respond to, as appropriate;

4. Clarifies that the financial institution's fair lending risk assessment should be conducted by independent actors within the institution or a third party;[115]
5. Explains the metrics and methods that the Agencies will use to evaluate compliance with fair lending laws;
6. Sets documentation and archiving requirements sufficient to ensure that financial institutions maintain the data, code, and information necessary for Agencies to review their AI/ML systems;
7. Sets explainability standards sufficient to enable the Agencies, advocates, consumers, independent auditors, and other key stakeholders to understand the decisions and outcomes generated by AI systems;
8. States that the Agencies will test for fair lending risk consistent with fair lending laws and policies, including by:
   a. Testing for disparate impact and less discriminatory alternatives;
   b. Ensuring that the training data is representative and accurate;
   c. Ensuring that the model measures lawful and meaningful attributes and seeks to predict valid target outcomes; and
   d. Ensuring that the technology is interpretable and its decision-making is sufficiently explainable to comply with fair lending laws;
9. To the maximum extent possible, ensures public access to detailed information about financial institutions' use of AI/ML and assessments of those models as well as the Agencies' reviews; and
10. Provides examples of best practices that financial institutions can use to mitigate fair lending risk.

- Public Research: The Agencies should encourage and support public research that analyzes the efficacy of specific uses of AI in financial services and the impact of AI in financial services for consumers of color and other protected classes. For example, the Agencies should encourage the CFPB and the Federal Housing Finance Agency to release more de-personalized loan-level data from the National Survey of Mortgage Originations and the National Mortgage Database so researchers, advocacy groups, and the public can study potential discriminatory and inequitable outcomes in the financial services market, especially as they relate to the use of AI.
  - For example, a research partnership could be formed between the National Institute of Standards and Technology ("NIST"), civil rights organizations, consumer protection

[115] This approach is consistent with the Model Risk Management Guidance, which states: "Validation involves a degree of independence from model development and use. Generally, validation should be done by people who are not responsible for development or use and do not have a stake in whether a model is determined to be valid." Model Risk Management Guidance at 9.

groups, non-profit research agencies, and financial institutions that rely on AI to evaluate how AI or machine learning models affect fair lending.[116]
    - The Agencies may also work with the National Science Foundation to ensure that a portion of the considerable allocations for research on AI focus on the implications of using AI in financial services.
    - The Agencies should find ways to include civil rights organizations with experience in working on fair lending issues on research projects.

- Specialized Fair Lending and AI Staff: The Agencies should immediately begin hiring staff with specialized skills that can provide guidance to financial institutions on assessing the impact of AI systems and that can review those assessments, particularly with respect to fair lending risks.
    - The Agencies should ensure that a financial institution's use of AI is reviewed by agency staff that specialize in AI and fair lending risks.
    - In addition, the Agencies should remind examination teams of the requirement under ECOA to examine for discriminatory impacts and less discriminatory alternatives, and to refer a matter to the U.S. Department of Justice if the agency has reason to believe that a creditor has engaged in a pattern or practice of discrimination.
    - Finally, given that AI models cannot be fully evaluated without considering the high risk of far-reaching and repeatable disparate impact, each Agency's fair lending and AI teams should be included in all meetings related to modeling efforts, including supervisory and enforcement issues, policy statements, and rulemakings.

- Transparency: The Agencies should prioritize transparency as they develop their understanding of the issues and proposed responses. First, the Agencies should strive to share their data, models, decisions, and proposed solutions so that all of the key stakeholders can stay apprised of and comment on the potential impact of proposed Agency actions on the national consumer financial market. Second, the Agencies should require financial institutions to share with the public as much information as possible regarding their AI systems and assessments of those systems to enable researchers and those impacted to evaluate the efficacy and impact of the systems.

- Engagement: The Agencies should stay engaged with a diverse group of key stakeholders, including civil rights organizations, consumer advocates, and impacted communities in order to receive ongoing input and feedback on these important decisions. The proposed solutions to AI risks are likely to have significant implications for borrowers and communities of color as well as other vulnerable communities, such as individuals with disabilities, families, and Limited English Proficiency borrowers. The Agencies should regularly engage with these communities and seek solutions that treat all borrowers and communities equitably.

[116] *See, e.g.*, NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software, NIST (Dec. 19, 2019).

Model-related Risk Mitigation Techniques

Please see the following responses for detailed information about model-related risk mitigation techniques:

- Overfitting at Question 6;
- Dynamic updating at Question 8; and
- Fair lending compliance at Question 11.

Non-quantitative Risk Mitigation Techniques

Risk mitigation should not be limited to model-related or quantitative methods, but should also include fair lending training; and diversity, equity, and inclusion.

- Fair Lending Training for All AI Stakeholders: The Agencies should ensure that all AI stakeholders—including regulators, financial institutions, and tech companies—receive regular fair lending and racial equity training. Trained professionals are better able to identify and recognize issues that may raise red flags. They are also better able to design AI systems that generate non-discriminatory and equitable outcomes. The more stakeholders in the field are educated about fair lending and equity issues, the more likely they are to create tools that expand opportunities for all consumers. Given the ever-evolving nature of AI, the training should be updated and provided on a periodic basis.

- Diversity, Equity, and Inclusion: The Agencies should ensure agency staff working on AI issues reflect diversity, including diversity based on race and national origin. In addition, the Agencies should encourage financial institutions to engage diverse staff for the AI development and design teams. Increasing the diversity of the regulatory and industry staff engaged in AI issues will lead to better outcomes for consumers. Research has shown that diverse teams are more innovative and productive[117] and that companies with more diversity are more profitable.[118] Moreover, people with diverse backgrounds and experiences bring unique and important perspectives to understanding how data impacts

[117] *See, e.g.,* John Rampton, Why You Need Diversity on Your Team, and 8 Ways to Build It, Entrepreneur (Sept. 6, 2019).
[118] *See, e.g.,* David Rock and Heidi Grant, Why Diverse Teams Are Smarter, Harvard Business Review (Nov. 4, 2016) (reporting that companies in the top quartile for ethnic and racial diversity in management were 35% more likely to have financial returns above their industry mean, and those in the top quartile for gender diversity were 15% more likely to have returns above the industry mean).

different segments of the market.[119] In several instances, it has been people of color who were able to identify potentially discriminatory AI systems.[120]

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?*

Effective model risk management principles and practices can aid fair lending evaluations of AI-based credit models. That said, the Agencies should expand these principles to ensure that "model risk" is defined to include discrimination risks, and they should instruct entities that internal fair lending assessments should be structured to ensure independence, effective challenge, and to guard against conflicts of interest.[121] Entities should separately be encouraged to consider periodic audits by independent third-parties.

Effective model risk management requires, among other things: a comprehensive system for cataloging, validating, and documenting model design, theory, and underlying logic; a rigorous assessment of data quality, comprehensiveness, and relevance; validation and documentation of model and variable performance; and monitoring use of models in production.[122]

Effective model risk management practices can aid compliance with fair lending laws. First, they can facilitate variable reviews by ensuring institutions understand the quality of data used and can identify potential issues, for example datasets that are over- or under-representative for certain populations. Second, they are essential to ensuring that models, and variables used within models, meet a legitimate business purpose by establishing that models meet performance standards to achieve the goals for which they were developed. Third, they establish a routine cadence for reviewing model performance. Fair lending reviews should, at a minimum, occur at the same cadence to ensure that models remain effective and are not causing new disparities because of, for example, demographic changes in applicant and borrower populations.

Fourth, model risk management principles should inform how fair lending reviews are conducted. For example, as part of effective model risk management, institutions will establish acceptable performance thresholds when comparing performance on development data to

---

[119] *See, e.g.,* Inioluwa Deborah Raji et al., Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing 39 (2020) (stressing the importance of "standpoint diversity" as algorithm development implicitly encodes developer assumptions of which they may not be aware). *See also* Model Risk Management Guidance at 4 (stating that "[a] guiding principle for managing model risk is 'effective challenge' of models, that is, critical analysis by objective, informed parties who can identify model limitations and assumptions and produce appropriate changes").

[120] *See, e.g.,* Steve Lohr, Facial Recognition is Accurate, if You're a White Guy, New York Times (Feb. 9, 2018) (explaining how Joy Buolamwini, a Black computer scientist, discovered that facial recognition worked well for her White friends but not for her).

[121] This approach is consistent with the Model Risk Management Guidance, which states: "Validation involves a degree of independence from model development and use. Generally, validation should be done by people who are not responsible for development or use and do not have a stake in whether a model is determined to be valid." Model Risk Management Guidance at 9. *See also* Christo Wilson et al., Building and Auditing Fair Algorithms: A Case Study in Candidate Screening (Mar. 2021).

[122] *See generally* Model Risk Management Guidance.

out-of-time or holdout data. These same thresholds can inform whether institutions should adopt less discriminatory alternatives. For example, at a minimum, an institution should adopt an alternative model if it is less discriminatory and any resulting drop in performance is within the thresholds the institution has set as acceptable for front-end model risk management purposes. Similarly, model risk management principles apply to validating models used by institutions that are developed by vendors or other third parties, even though some processes may be modified in accordance with guidance from regulators.[123] Applying these same principles to fair lending reviews requires that institutions review models developed by vendors for compliance with fair lending laws, even though modifications to the institution's normal process for conducting these reviews may be required.

Despite the important role that model risk management can play in supporting fair lending compliance, the Agencies have yet to issue a policy statement that explicitly articulates this message and that tailors model risk management principles to the fair lending compliance context. The Federal Reserve Board, OCC, and FDIC recently issued an Interagency Statement on Model Risk Management for Bank Systems Supporting Bank Secrecy Act ("BSA")/Anti-Money Laundering ("AML") Compliance.[124] The statement is intended to clarify how the previously-released model risk management guidance may be a useful resource to guide a bank's model risk management framework with respect to BSA/AML compliance. Similarly, the Agencies should issue a statement clarifying how the model risk management guidance applies with respect to fair lending compliance.

Recommendations

Accordingly, our organizations recommend the following:

- Policy Statement regarding Model Risk Management and Fair Lending Compliance: The Agencies should issue a joint policy statement that clearly explains how the principles of model risk management relate to systems or models used by financial institutions to assist in complying with ECOA, Regulation B, and the Fair Housing Act. Moreover, this policy statement should define "model risk" to include the risk of discriminatory or inequitable outcomes for consumers, rather than just the risk of financial loss to a financial institution. The Agencies should instruct entities on how to ensure that internal assessments are performed under conditions that maximize independence and effective challenge, and that guard against conflicts of interests. Where appropriate, financial institutions should also be encouraged to separately engage third-party auditors to assess model discrimination risks.

14. *As part of their compliance management systems, financial institutions may conduct fair lending risk assessments by using models designed to evaluate fair lending risks ("fair lending risk assessment models"). What challenges, if any, do financial institutions face*

[123] *See, e.g.*, FDIC, FIL-22-2017, Supervisory Guidance on Model Risk Management (June 7, 2017).
[124] Federal Reserve Board, OCC, FDIC, Interagency Statement on Model Risk Management for Bank Systems Supporting Bank Secrecy Act/Anti-Money Laundering Compliance (Apr. 9, 2021).

when applying internal model risk management principles and practices to the development, validation, or use of fair lending risk assessment models based on AI?

Institutions' fair lending risk assessments should include routine reviews of models for potential disparate treatment and disparate impact, including assessing whether legitimate business needs can be met with less discriminatory alternative models. Many institutions conduct disparate impact and alternatives analyses on credit-related models, both before production and on a regular cadence for models in production. For example, these institutions routinely evaluate their credit-related models for disparate impact risk and, to the extent models have a discriminatory effect, they actively search for alternatives that maintain performance while minimizing impact.

While some lenders have institutionalized disparate impact fair lending testing protocols, others have not. Some institutions have suggested they are reluctant to conduct such tests, especially on models that are in use and that did not undergo disparate impact testing before production, for fear that they might be creating legal risk. Unfortunately, this position shows a clear weakness in the institution's Compliance Management System and has the counterproductive effect of potentially discriminatory models remaining in use, even if less discriminatory alternatives may exist. In the absence of a robust fair lending compliance framework, these entities may end up violating the fair lending laws, which will perpetuate discrimination and structural inequality.

Recommendations

Accordingly, our organizations recommend the following:

- Supervisory Issues: The Agencies should ensure that financial institutions' use of AI is reviewed by staff that specialize in AI and fair lending risk.

- Policy Statement regarding Supervisory Expectations and Methodologies: The Agencies should explain in detail their expectations and methodologies for evaluating a financial institution's fair lending risk assessment models based on AI. In addition to the recommendations provided here, please see the recommendations in the response to Question 12.
    - In particular, the Agencies should confirm that the financial institution's self-assessment of compliance with federal consumer financial law must include disparate impact and less discriminatory alternatives analyses.
    - Also, the Agencies should make clear that, consistent with the Uniform Interagency Consumer Compliance Rating System[125] and guidance on responsible business conduct,[126] the existence and robustness of AI-based fair lending risk assessments will be considered, along with other relevant factors, in addressing violations of federal consumer financial law in supervisory and enforcement matters.

[125] FFIEC, Uniform Interagency Consumer Compliance Rating System (Nov. 7, 2016).
[126] CFPB Bulletin 2020-01, Responsible Business Conduct: Self-Assessing, Self-Reporting, Remediating, and Cooperating (Mar. 6, 2020).

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.*

Please see the response to Questions 1-3 regarding Explainability.

F. Additional Considerations

16. *To the extent not already discussed, please identify any additional uses of AI by financial institutions and any risk management challenges or other factors that may impede adoption and use of AI.*
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.*

Please see the Background and Global Recommendations section of this response.

Thank you for considering our views.

Sincerely,

AI Blindspot
American Civil Liberties Union
Americans for Financial Reform Education Fund
California Reinvestment Coalition
Center for Democracy & Technology
Center for New York City Neighborhoods, Inc.
Center for Responsible Lending
Center on Race, Inequality, and the Law
Consumer Action
Consumer Federation of America
East Metro Civic Alliance
Equal Rights Center
Fair Housing Advocates Association
Fair Housing Advocates of Northern California
Fair Housing Center of Central Indiana
Fair Housing Center of Northern Alabama
Fair Housing Center of Southwest Michigan
Fair Housing Center of West Michigan

Fair Housing Council of Greater San Antonio
FairPlay AI
Housing Equality Center of Pennsylvania
Illinois People's Action
Lawyers' Committee for Civil Rights under Law
The Leadership Conference on Civil and Human Rights
Liberation in a Generation
Long Island Housing Services, Inc.
Louisiana Fair Housing Action Center
Miami Valley Fair Housing Center
NAACP Legal Defense and Educational Fund, Inc. (LDF)
National Community Reinvestment Coalition
National Consumer Law Center (on behalf of its low-income clients)
National Council of Asian Pacific Americans
National Fair Housing Alliance
New York University, Center for Critical Race & Digital Studies
North Texas Fair Housing Center
Philadelphia Unemployment Project
SolasAI
South Suburban Housing Center
Student Borrower Protection Center
TechEquity Collaborative
Texas Appleseed
Upturn
U.S. PIRG
Woodstock Institute