September 10, 2021

Dr. James Olthoff, Director
National Institute of Standards and Technology
Attn: Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899-2000
ai-bias@list.nist.gov

Re: Draft NIST Special Publication 1270 – A Proposal for Identifying and Managing Bias within Artificial Intelligence

Dear Director Olthoff,

We the undersigned civil rights, consumer, technology, and other advocacy organizations are writing in response to the National Institute of Standards and Technology's ("NIST") request for comment on NIST's Proposal for Identifying and Managing Bias within Artificial Intelligence ("NIST Proposal").[1]

We applaud NIST for seeking input on this important topic of bias and discrimination in the context of artificial intelligence ("AI"). Our organizations believe that the responses below will help inform NIST's policies for identifying, analyzing, and managing bias in AI systems.[2]

I. The NIST Proposal Offers an Initial Framework for Identifying and Managing AI Bias

In this Proposal, NIST concludes that:

- Bias is neither new nor unique to AI.
- The goal is not zero risk but rather, identifying, understanding, measuring, managing, and reducing bias.
- Standards and guides are needed for terminology, measurement, and evaluation of bias.
- Bias reduction techniques that are flexible and that can be applied across contexts, regardless of industry, are needed.

[1] NIST, A Proposal for Identifying and Managing Bias within Artificial Intelligence, Draft NIST Special Publication 1270 (June 22, 2021).

[2] Note on the language used in this comment letter: 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.

Based on these conclusions, NIST proposes a three-stage approach for analyzing and managing AI bias:

1. Pre-Design: where the technology is devised, defined and elaborated;
2. Design and Development: where the technology is constructed; and
3. Deployment: where technology is used by, or applied to, various individuals or groups.

NIST's approach is intended to foster discussion about the path forward and collaborative development of standards and a risk-based framework. Accordingly, NIST plans to:

- Develop a framework for trustworthy and responsible AI with the participation of a broad set of stakeholders to ensure that standards and practices reflect viewpoints not traditionally included in AI development; and
- Collaboratively develop additional guidance for assurance, governance, and practice improvements as well as techniques for enhancing communication among different stakeholder groups.

We commend NIST for its work to date and offer the following background and recommendations to enhance the Proposal. Although the majority of our comments are focused on discrimination risks in the context of housing and consumer credit, the observations and recommendations can be applied across contexts and industries to inform the broader principles of analyzing and managing AI bias. We also recommend that NIST review the advocate response to the federal financial regulators' Request for Information and Comment on Financial Institutions' Use of Artificial Intelligence, including Machine Learning. [3]

II. 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")[4] 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.[5] Notably, the data used to create the maps were not just collected

[3] NFHA, Leading Civil Rights, Consumer, and Technology Advocates Urge the Federal Financial Regulators to Promote Equitable Artificial Intelligence in Financial Services (July 1, 2021).

[4] 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.*

[5] *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).

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.[6] 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.[7]

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 Federal Housing Administration, and other entities that perpetuated discriminatory practices. Government, industry, and 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.

III. Existing Civil Rights Laws and Policies Provide a Framework for Identifying, Analyzing, and Addressing the Risk of Discrimination in AI

Various civil rights laws prohibit discrimination in contracts, housing, credit, employment, and other critical areas. For example, two primary federal anti-discrimination laws—the Equal Credit Opportunity Act ("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.[8] 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

---

[6] *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).
[7] *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).
[8] 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.

estate-related transactions.[9] In addition, the federal financial regulators[10] have issued several policies that provide a framework for risk identification, analysis, and management, 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.[11]

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 generally 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.[12] These frameworks translate well to the identification, analysis, and mitigation of discrimination risk in AIs, 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 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.[13]

---

[9] Other laws, such as UDA(A)P authorities, can also be applied to prevent discrimination in consumer financial services, although to date regulators 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).

[10] The "federal financial regulators" include the Board of Governors of the Federal Reserve ("Federal Reserve Board"), Consumer Financial Protection Bureau ("CFPB"), Federal Deposit Insurance Corporation ("FDIC"), National Credit Union Administration ("NCUA"), and Office of the Comptroller of the Currency ("OCC").

[11] Federal Financial Institutions Examination Council ("FFIEC"), Revised FFIEC Fair Lending Examination Procedures and Use of Specialized Examination Techniques, (Aug. 4, 2009); Federal Reserve Board and OCC, Supervisory Guidance on Model Risk Management, SR 11-7 at 3 (Apr. 4, 2011) ("Model Risk Management Guidance"); 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).

[12] *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). *See also*, U.S. Department of Housing and Urban Development ("HUD"), Reinstatement of HUD's Discriminatory Effects Standard, 86 Fed. Reg. 33590 (June 25, 2021).

[13] *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).

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 policies provide a framework for identifying, analzying, and managing the risk of discrimination in AI. That said, NIST and the other federal agencies should coordinate to ensure consistent compliance and fair outcomes by, among other things, clarifying certain ambiguities that relate to AI, setting robust regulatory expectations regarding testing for AI, and ensuring models are non-discriminatory and equitable, as discussed below.

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

In some respects, the U.S. is behind in advancing non-discriminatory and equitable technology. 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 for consumers who are harmed by current practices—and to rethink appropriate guardrails to promote a safe, fair, and inclusive market. NIST and the other federal agencies have an opportunity to rethink comprehensively how they regulate decisions that determine who has access to opportunities and on what terms.

As NIST and the other federal agencies consider their approach to the use of AI, the European Union's recently-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 industry) and how to tier risk based on the intended use of the AI system.[14] 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.[15] Importantly, the EU made this determination based on explicit recognition of (i) the importance of credit evaluations to fully

[14] 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. regulators as a significant set of American firms may be already complying with the EU's framework.

[15] *Id*. at Annex III.

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 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**.[16]

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 Certain State Laws Provide a Useful Starting Point for a Robust 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 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.[17]

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.[18] Finally, providers of AI systems must ensure that natural persons are informed that they are interacting with an AI system.[19] Penalties by regulators for non-compliance would be as high as 6% of the entity's total global earnings (before costs).[20]

The U.S. should further improve on the EU proposal by ensuring transparency and regulatory review of provider self assessments.[21] Federal agencies should require that regulated entities

[16] *Id*. at Recital 37 (emphasis added).
[17] *Id*. at Titles III and VIII.
[18] *Id*. at Title III, Ch. 3 and 5.
[19] *Id*. at Title IV, Art. 52.
[20] *Id*. at Title X, Art. 71.
[21] *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

conduct discrimination risk assessments that detail how their AI 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. Federal agencies should require that regulated entities 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. 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, NIST 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.[22]

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."[23]

Intelligence, NewYork Times (Apr. 21, 2021) (quoting an advocacy group that is critical of the Proposed Regulation's reliance on self assessments).

[22] 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).

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

V. Data Quality Can Be a Significant Risk at the Pre-Design Stage

In the Proposal, NIST appropriately focused on the pre-design stage as an important stage in the AI lifecycle and in managing the risk of discrimination. The review of data quality at the pre-design stage is critically important to the development of an AI model that is non-discriminatory, equitable, and effective for its intended purpose.[24] One of the core problems in AI is sometimes referred to as "garbage in, garbage out." That is, modern AI algorithms are "trained" on the basis of what's happened in the past. If the data on which an AI system 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. Following are some examples to supplement NIST's thinking regarding data quality risks at the pre-design stage.

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 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.[25]

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.[26] 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. With limited explicit exceptions, it is a violation of

---

[24] *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").

[25] *See, e.g*, Kenneth P. Brevoort, Phillipp Grimm, and Michelle Kambara, Data Point: Credit Invisibles, CFPB Office of Research, 6 (May 2015). *See also* Will Douglas Heaven, Bias Isn't the Only Problem with Credit Scores - and, No, AI Can't Help, MIT Technology Review (June 17, 2021); Laura Blattner and Scott Nelson, How Costly Is Noise? Data and Disparities in Consumer Credit (May 17, 2021).

[26] 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).

the fair lending laws' prohibitions against overt, intentional discrimination to use a protected class as a variable in a credit scoring or pricing model.[27] This is equally true for close proxies, such as zip code, geographic location, or language preference.[28] 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.

Alternative Data: Traditional credit history scores reflect immense racial disparities due to extensive historical and ongoing discrimination.[29] Black and Latino consumers are less likely to have credit scores in the first place, limiting their access to financial services.[30] There is an obvious need for better, fairer, and more inclusive measures of creditworthiness.[31] 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.[32] 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.[33] 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."[34]

---

[27] *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).

[28] *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.").

[29] *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).

[30] *See* Kenneth P. Brevoort, Philipp Grimm, and Michelle Kambara, Data Point: Credit Invisibles, CFPB (May 2015).

[31] *See* Chi Chi Wu, Reparations, Race, and Reputation in Credit: Rethinking the Relationship between Credit Scores and Reports with Black Communities (Aug. 7, 2020).

[32] *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).

[33] *See* FinRegLab, The Use of Cash-Flow Data in Underwriting Credit (July 2019).

[34] *See* Federal Reserve Board, CFPB, FDIC, OCC, NCUA, Interagency Statement on the Use of Alternative Data in Credit Underwriting (Dec. 3, 2019).

VI. It Can Be Critically Important to Review Model Risks at the Design/Development and Deployment Stages

In the Proposal, NIST appropriately focused on two other important stages in the AI lifecycle: (1) design and development, and (2) deployment. These in-process and post-process reviews can be critically important to the development of an AI model that is non-discriminatory, equitable, and effective for its intended purpose. AI models can increase risk due to the models' greater complexity and their potential to exacerbate historical disparities and flaws in underlying data. Following are some examples to supplement NIST's thinking regarding model risks at the design/development and deployment stages.

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.[35] 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.[36] 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.[37] 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.[38] Similarly, predatory lending algorithms could result in digital reverse

---

[35] *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).
[36] *See* National Fair Housing Alliance, Facebook Settlement (Mar. 19, 2019).
[37] LaTanya Sweeney, Discrimination in Online Ad Delivery, 11(3) ACMQueue (Apr. 2, 2013).
[38] *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.[39]

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.[40] 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.[41] As noted, both ECOA and the Fair Housing Act prohibit disparate impact in certain types of credit.[42] Consistent with existing law and policy, AI systems related to consumer credit 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.[43]

VII. NIST Should Adopt These Recommendations to Enhance Its Proposal

These recommendations focus on the responsibility of NIST to propose a framework that ensures that AI risk identification, analysis, and mitigation work hand in hand with discrimination risk identification, analysis, and management, and that efforts to identify, analyze, and manage AI risks do not exclude or undermine efforts to promote fair and equitable outcomes. In particular, any new AI-related initiatives should be reviewed for 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 to enhance the NIST Proposal and the NIST process.

---

[39] *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).
[40] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace, Consumer-Lending Discrimination in the FinTech Era, UC Berkeley (2019).
[41] *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").
[42] *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).
[43] *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).

Enhancements to NIST's Proposal:

- Non-discrimination and Equity: NIST should recommend that practitioners take the steps needed to ensure non-discriminatory and equitable outcomes for all end users of AI systems. Most importantly, NIST should ensure that "model risk" is defined to include the risk of discriminatory or inequitable outcomes for consumers, rather than just the risk of financial loss to an entity.[44] That is, the analysis of discrimination risk and equity should be integrated into all AI discussions and not treated as an afterthought or separate issue.

- Actionable Policies: Existing civil rights laws and policies provide a framework for NIST and the other federal agencies to identify, analyze, and manage discrimination risk in AI. That said, NIST and the other federal agencies can be more effective in ensuring consistent compliance and fair outcomes by setting clear and robust expectations regarding testing and ensuring models are non-discriminatory and equitable. NIST and the other federal agencies have been in learning mode for some time, which may have put the U.S. behind in advancing non-discriminatory and equitable technology. To retain our competitive edge in global society, NIST and the other federal agencies should move quickly to issue actionable policy statements that clearly state their commitment to consumer protection and civil rights laws, including fair lending laws; provide insight into their expectations and methods; and provide useful guardrails and best practices. The time to act is now as the use of AI proliferates in every aspect of society and has the potential for far-reaching adverse impacts for people and communities of color and other protected groups. More specifically, NIST can be more effective in ensuring consistent compliance and fair outcomes by moving quickly to include in their framework 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 entities should be aware of and control for;
    3. Sets clear standards for discrimination risk assessments, including:
        a. How to conduct discrimination testing and evaluation throughout the AI model's conception, design, implementation, and use; and
        b. What information must be detailed in the documentation of discrimination risk assessments, including:
            (i) What testing has been conducted and less discriminatory alternatives have been considered; and
            (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.
        c. Clarification that the discrimination risk assessment should be conducted by independent actors within the institution or a third party;[45]

[44] *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").

[45] 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

4. Sets documentation and archiving requirements sufficient to ensure that entities maintain the data, code, and information necessary to review AI systems;
5. Sets explainability standards sufficient to enable regulators, advocates, consumers, independent auditors, and other key stakeholders to understand the decisions and outcomes generated by AI systems;
6. States that, where applicable, regulators will test for discrimination risk consistent with civil rights 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 civil rights laws;
7. To the maximum extent possible, ensures public access to detailed information about an entity's use of AI and assessments of those models as well as regulatory reviews; and
8. Provides examples of best practices that entities can use to mitigate discrimination risk.

- Diversity, Equity, and Inclusion: NIST should encourage entities engaged in AI model development and deployment to ensure staff working on AI issues reflect diversity, including diversity based on race and national origin. Increasing staff diversity will lead to better outcomes for consumers and other affected parties. Research has shown that diverse teams are more innovative and productive[46] and that companies with more diversity are more profitable.[47] Moreover, people with diverse backgrounds and experiences bring unique and important perspectives to understanding how data impacts different segments of the market.[48] In several instances, it has been people of color who were able to identify potentially discriminatory AI systems.[49]

- Civil Rights Training for All AI Stakeholders: NIST should encourage all AI stakeholders to receive regular civil rights and racial equity training. Trained

---

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.

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

[47] 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).

[48] 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").

[49] 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).

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 discrimination and equity issues, the more likely they are to create tools that expand opportunities for all. Given the ever-evolving nature of AI, the training should be updated and provided on a periodic basis.

- Transparency for AI Providers: The NIST proposal should recommend that AI providers 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.

Enhancements to NIST's Process:

- Action Plan: After review of the comments, NIST should immediately issue a detailed Action Plan, which may include plans for a white paper, policy statement, or a proposed regulation.

- Engagement: NIST 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 people and communities of color as well as other vulnerable communities, such as individuals with disabilities, families, and Limited English Proficiency consumers. NIST should regularly engage with these communities and seek solutions that treat all people and communities equitably.

- Specialized Civil Rights Staff: NIST should immediately begin hiring staff with specialized skills that can provide guidance to entities on assessing the potentially discriminatory impact of AI systems and that can review those assessments, particularly with respect to discrimination risks.

- Diversity, Equity, and Inclusion at NIST: For the reasons noted above, NIST should ensure agency staff working on AI issues reflect diversity, including diversity based on race and national origin.

- Civil Rights Training for NIST Staff: For the reasons noted above, NIST should ensure that agency staff receive regular civil rights training.

- Transparency for NIST: NIST should prioritize transparency as it develops its understanding of the issues and proposed solutions. NIST should strive to share its methodology, 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 actions.

- Public Research: NIST should encourage and support public research that analyzes the efficacy of specific uses of AI and the impact of AI for people and communities of color and other protected classes. For example, a research partnership could be formed between 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 or other aspects of civil rights.[50]

Thank you for considering our views. If you have any questions, please contact Michael Akinwumi, Chief Tech Equity Officer (makinwumi@nationalfairhousing.org), or Maureen Yap, Senior Counsel (myap@nationalfairhousing.org), National Fair Housing Alliance.

Sincerely,

AI Blindspot

Americans for Financial Reform Education Fund

California Reinvestment Coalition

Center for Community Progress

Center for Responsible Lending

Consumer Action

Consumer Federation of America

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

Integrated Community Solutions, Inc.

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

The Leadership Conference on Civil and Human Rights

Long Island Housing Services, Inc.

Louisiana Fair Housing Action Center

Miami Valley Fair Housing Center, Inc.

MICAH- Metropolitan Interfaith Council on Affordable Housing

Mountain State Justice

NAACP Legal Defense and Educational Fund, Inc. (LDF)

National CAPACD

National Coalition For The Homeless

National Community Reinvestment Coalition

National Consumer Law Center (on behalf of its low-income clients)

National Fair Housing Alliance

North Texas Fair Housing Center

NYU Center on Race, Inequality, and the Law

SolasAI

South Suburban Housing Center

Southern Poverty Law Center Action Fund

Woodstock Institute