October 18, 2019

SUBMITTED VIA REGULATIONS.GOV

Office of the General Counsel
Rules Docket Clerk
Department of Housing and Urban Development
451 Seventh Street SW, Room 10276
Washington, DC 20410-0001

Re: Reconsideration of HUD's Implementation of the Fair Housing Act's Disparate Impact Standard, Docket No. FR-6111-P-02

Dear Sir or Madam,

The undersigned individuals and organizations represent expertise in the fields of computer science, statistics, and digital and civil rights. We write to offer comments in response to the above-docketed notice of proposed rulemaking ("NPRM") concerning proposed changes to the disparate impact standard as interpreted by the U.S. Department of Housing and Urban Development ("HUD"). The NPRM creates new pleading hurdles that would make it practically impossible for plaintiffs to have their disparate-impact cases heard, especially when the alleged discrimination results from algorithmic models. The NPRM's algorithmic defenses seriously undermine HUD's ability to address discrimination, are unjustified in the record, and have no basis in computer or data science. Adopting the NPRM would violate HUD's obligation to end discriminatory housing practices and "affirmatively further fair housing."[1]

Algorithms are being used to make decisions that impact the availability and cost of housing. These decisions include screening rental applicants,[2] underwriting mortgages,[3] determining the cost of insurance,[4] and targeting online housing offers.[5] These models are seldom designed to take protected

[1] 42 U.S.C § 3608(d); 78 Fed. Reg. 11460, 11465, 11477 (2013).

[2] *See, e.g.*, TransUnion SmartMove, https://www.mysmartmove.com/ (last visited Oct. 17, 2019); appfolio Property Manager, Built-in Tenant Screening, https://www.appfolio.com/features/residents-and-leasing#tenant-screening (last visited Oct. 17, 2019).

[3] *See, e.g.*, Fannie Mae Desktop Underwriter, https://www.fanniemae.com/singlefamily/desktop-underwriter (last visited Oct. 17, 2019).

[4] *See* Robert Bartlett et al., *Consumer Lending Discrimination in the FinTech Era* 1, https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf ("Consumer lending in the United States is changing rapidly, with loan origination becoming almost exclusively algorithmic. A case in point is the Rocket Mortgage of the platform lender Quicken, which is the largest-volume mortgage product in the U.S. as of 2018.").

[5] *See, e.g.*, Sheryl Sandberg, *Doing More to Protect Against Discrimination in Housing, Employment and Credit Advertising*, Facebook Newsroom (Mar. 19, 2019), https://newsroom.fb.com/news/2019/03/protecting-against-discrimination-in-ads/; Tracy Jan & Elizabeth Dwoskin, *Facebook Agrees to Overhaul Targeted Advertising System for Job, Housing and Loan Ads After Discrimination Complaints* (Mar. 19, 2019), https://www.washingtonpost.com/business/economy/facebook-agrees-to-dismantle-targeted-advertising-system-

characteristics into account, yet they still have the capacity for protected-class discrimination.[6] The datasets and correlations on which they rely can reflect societal bias[7] in non-obvious ways that models may reproduce or reinforce.[8] Models may also create or mask discrimination and bias without regard to societal bias in data.[9] This is precisely the type of discrimination that disparate-impact liability is supposed to address.

The Fair Housing Act ("FHA") prohibits not only intentional discrimination but also practices that result in discriminatory effects regardless of intent.[10] HUD and the courts have recognized disparate-impact liability for over 25 years.[11] Under HUD's existing disparate impact rule, when a plaintiff makes a prima facie showing of disparate impact, the burden of proof shifts to the defendant to show a business justification for the alleged practices, and the plaintiff can respond by proving that less discriminatory alternatives exist.[12] This "burden-shifting framework" allows courts to examine the specific facts of the case and determine whether the facially neutral practice caused unjustified discrimination. Under the

for-job-housing-and-loan-ads-after-discrimination-complaints/2019/03/19/7dc9b5fa-4983-11e9-b79a-961983b7e0cd_story.html.

[6] *See, e.g.*, Bartlett et al., *supra* note 4, at 16 (finding that FinTech algorithms discriminated 40% less than face-to-face mortgage lenders but still charged Latinx and African-American applicants 5.3 basis points more in interest for purchase mortgages and 2.0 basis points more for refinance mortgages originated on FinTech platforms) ("[A]lgorithmic lending may reduce discrimination relative to face-to-face lenders, but algorithmic lending is not alone sufficient to eliminate discrimination in loan pricing."); Ziad Obermeyer & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million People*, Proceedings of the Conference on Fairness, Accountability, and Transparency 89-89 (2019), https://dl.acm.org/citation.cfm?id=3287593 (finding that an algorithm used to enroll patients in a care management program underestimated the health risk and thus under-enrolled black patients compared to their white counterparts, because of the model designers' decision to focus on cost rather than other factors).

[7] Housing-related decisions and data inz the United States are inevitably shaped by centuries of segregationist policies and institutional discrimination. *See, e.g.*, Joseph D. Rich, Lawyers' Committee for Civil Rights Under Law, HUD's New Discriminatory Effects Regulation: Adding Strength and Clarity to Efforts to End Residential Segregation (May 2013), https://lawyerscommittee.org/wp-content/uploads/2015/08/HUDs-New-Discriminatory-Effects-Regulation.pdf (citing Trafficante v. Metro. Life Ins. Co., 409 U.S. 205, 211 (1972)) ("Our nation's highly segregated housing patterns did not occur by accident. Rather, it is well-established that they are a product of a complex web of private and public policies, practices and decisions made since the beginning of the 20th century. Impediments to addressing residential segregation have been a core concern of the FHA since its passage because such impediments hinder advancement of the FHA's goal of achieving 'truly integrated and balanced living patterns.'").

[8] *See, e.g.*, Jieyu Zhao et al., *Men Also Like Shopping: Reducing Gender Bias Amplification Using Corpus-Level Constraints*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (2017), https://arxiv.org/abs/1707.09457 (finding that models trained on datasets containing gender bias amplified that bias).

[9] *See, e.g.*, Obermeyer & Sendhil, *supra* note 6 ("The root cause of this bias is not . . . the underlying data, but the algorithm's objective function itself.").

[10] 24 C.F.R. § 100.500; Texas Dep't of Hous. & Cmty. Affairs v. Inclusive Cmtys. Proj., Inc., 135 S. Ct. 2507 (2015).

[11] *See* 78 Fed. Reg. 11460, 11462 (2013) (citing Memorandum from the HUD Assistant Secretary for Fair Housing & Equal Opportunity, The Applicability of Disparate Impact Analysis to Fair Housing Cases (Dec. 17, 1993)).

[12] 24 C.F.R. § 100.500.

NPRM, most algorithmic models that caused cognizable disparate impacts would never make it to this fact-specific inquiry.[13]

Under the NPRM, a defendant relying on an algorithmic model could defeat a claim at the prima facie stage by showing that (a) the model's inputs do not include close proxies for protected classes; (b) a neutral third-party determined that the model has predictive value; or (c) a third party created the model. These defenses do nothing to disprove discrimination and undermine efforts to address it. There is no method for restricting a model's inputs that can guarantee that an algorithm will not produce discriminatory outcomes, nor are there common industry standards for preventing algorithmic discrimination. In many cases, researchers have discovered bias *only* after testing the models' outcomes.[14] Indeed, the type of discrimination that algorithmic models create is precisely the type of discrimination that the existing disparate impact test was designed to uncover.

We oppose the NPRM because it would unjustifiably shield users of algorithmic models from liability under the FHA. When a plaintiff alleges a disparate impact based on an algorithmic model, courts should analyze the model under HUD's existing burden-shifting framework, which allows for the discovery and fact finding necessary to determine whether a model causes unjustified discriminatory effects and whether less discriminatory alternatives exist.

I. The proposed rule would shield actors who use discriminatory algorithmic models from liability without justification

Without justification or explanation, the proposed rule would create three defenses that could be used to dismiss a disparate-impact claim based on an algorithmic model at the prima facie stage.[15] If a plaintiff alleges that "the cause of a discriminatory effect is a model . . . such as a risk assessment algorithm," the defendant can have the claim dismissed by:

> (i) Provid[ing] the material factors that make up the inputs used in the challenged model and show[ing] that these factors do not rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act and that the model is predictive of credit risk or other similar valid objective;

[13] *See*, *e.g.* Morgan Williams, National Fair Housing Alliance, HUD's Proposed Rule is a Direct Assault on the Supreme Court's Decision in Inclusive Communities (Oct. 10, 2019), https://nationalfairhousing.org/2019/10/10/huds-proposed-rule-is-a-direct-assault-on-the-supreme-courts-decision-in-inclusive-communities/ (describing how the NPRM would make it impossible to bring prototypical "heartland" disparate impact cases).

[14] *See, e.g.*, Bartlett et al., *supra* note 4; Muhammad Ali et al., *Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Skewed Outcomes*, Proceedings of the ACM on Human-Computer Interaction (2019), https://arxiv.org/abs/1904.02095.

[15] 84 Fed. Reg. 42854, 42862 (2019).

> (ii) Show[ing] that the challenged model is produced, maintained, or distributed by a recognized third party that determines industry standards, the inputs and methods within the model are not determined by the defendant, and the defendant is using the model as intended by the third party; or
>
> (iii) Show[ing] that the model has been subjected to critical review and has been validated by an objective and unbiased neutral third party that has analyzed the challenged model and found that the model was empirically derived and is a demonstrably and statistically sound algorithm that accurately predicts risk or other valid objectives, and that none of the factors used in the algorithm rely in any material part on factors that are substitutes or close proxies for protected classes under the Fair Housing Act.[16]

These defenses would do little to show that an algorithmic model is not discriminatory. There is no method for restricting possible inputs or set of industry standards that can guarantee that a model will not be discriminatory. Assessing disparate impact in algorithmic models requires testing or foreseeing the models' outcomes in the context in which it is used. Algorithmic models raise precisely the risk of discrimination that disparate impact is supposed to address. As this comment will discuss, models can reflect and perpetuate discriminatory patterns even without using obvious proxies for protected classes or intentionally discriminatory design. If these risks are not mitigated, models can be used to make systematically discriminatory decisions at scale by codifying discrimination into the rules that govern the housing market.

A. Removing or omitting "close proxies" for protected classes is inadequate to protect against discrimination.

The first and third defenses would allow a defendant to defeat a prima facie claim of disparate impact by showing (or having a neutral third party determine) that "the inputs used in the challenged model . . . do not rely . . . on factors that are substitutes or close proxies for protected classes."[17] This is scientifically incoherent. It suggests that if a model's individual components are not close correlates with protected classes then the model as a whole will not cause disparate impacts. This assumption is false and is contradicted by the relevant peer-reviewed research.[18]

Machine-learning models' predictions do not depend on any one input in isolation. They depend on relationships between inputs. Factors that are not close proxies for protected class in isolation may become close proxies when combined.[19] For example, surname and zip code are both weak proxies for

[16] *Id.*

[17] *Id.*

[18] *See* Amanda Bower et al., *Fair Pipelines* (July 3, 2017), https://arxiv.org/abs/1707.00391; Cynthia Dwork & Christina Ilvento, *Fairness Under Composition* (2018), https://arxiv.org/abs/1806.06122.

[19] *See, e.g.*, Samuel Yeom, Anupam Datta & Matt Fredrikson, *Hunting for Discriminatory Proxies in Linear Regression Models*, Proceedings of the 32nd Conference on Neural Information Processing Systems (2018), https://www.cs.cmu.edu/~mfredrik/papers/Yeom18-nips.pdf (finding that many variables combined in a

race in isolation, but when combined they can be stronger proxies.[20] In online advertising markets where advertisers select interest-based categories such as "frequent travelers" and "online games" to help target their ads to people with those interests, multiple interest categories combined may be a much stronger predictor of protected class than any one interest category in isolation.[21]

The same input may have differential predictive value across population groups. For example, if one group of high school seniors hires an SAT tutor and takes the SAT multiple times, reporting the highest score, and another group does not hire a tutor and takes the test only once, the relationship between SAT score and performance in college will presumably be different for each group.[22] A model designed to predict college performance based on SAT score for all students will generally have higher predictive value for the majority population and lower predictive value for minority populations.[23]

Because most algorithmic models are used by humans to help make decisions, even a model whose outputs achieve statistical parity (or some other fairness measure) across different groups can lead to discriminatory decisionmaking by a human in the loop.[24] For example, a leasing agency that uses an algorithm to screen applicants' risk might give less weight to negative risk scores for applicants who were referred by existing tenants than for all other applicants. The agency's decisions about which applicants to accept could result in an unlawful disparate impact even if the screening tool itself does not create disparate risk scores.

There is no precedent justifying HUD's reliance on inputs alone to dispel an allegation of disparate impact. The Fair and Accurate Credit Transactions Act of 2003 (FACT Act) ordered the Federal Reserve Board (Fed) and the Federal Trade Commission (FTC), in consultation with HUD, to conduct a study of

predictive policing model together acted as a strong proxy for race but that no individual variable was a particularly strong proxy for race, thus concluding that "in practice multiple variables combine to result in a stronger proxy than any of the individual variables"); David Skanderson & Dubravka Ritter, Payment Cards Center Discussion Paper, Fair Lending Analysis of Credit Cards 28, https://www.philadelphiafed.org/-/media/consumer-finance-institute/payment-cards-center/publications/discussion-papers/2014/D-2014-Fair-Lending.pdf.

[20] *See* Consumer Financial Protection Bureau, Using Publicly Available Information to Proxy for Unidentified Race & Ethnicity: A Methodology & Assessment (2014), https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf. *But see* Jiahao Chen et al., *Fairness Under Unawareness: Assessing Disparity When Protected Class is Unobserved*, Conference on Fairness, Accountability & Transparency (FAT*) (2019), https://arxiv.org/pdf/1811.11154.pdf.

[21] Till Speicher et al., *Potential for Discrimination in Online Targeted Advertising*, Proceedings of Machine Learning Research 81:1, 8–12 (2018), http://proceedings.mlr.press/v81/speicher18a/speicher18a.pdf.

[22] *See* Alexandra Chouldechova & Aaron Roth, *The Frontiers of Fairness in Machine Learning* (Oct. 23, 2018), https://arxiv.org/pdf/1810.08810.pdf.

[23] *See Id.* at 2.

[24] A phenomenon called "automation bias" can lead human decision-makers to give undue weight to a computer generated prediction. *See, e.g.*, Kathleen L. Mosier et al., *Automation Bias: Decision Making & Performance in High-Tech Cockpits*, 8 Internat'l J. of Aviation Psych. 47 (1998), https://www.researchgate.net/profile/Linda_Skitka/publication/11805395_Automation_Bias_Decision_Making_and_Performance_in_High-Tech_Cockpits/links/0912f511ea2be025b8000000/Automation-Bias-Decision-Making-and-Performance-in-High-Tech-Cockpits.pdf; Raja Parasuraman & Victor Riley, *Humans and Automation: Use, Misuse, Disuse, Abuse*, 39 Hum. Factors 230 (1997).

the effects of credit scoring, including negative or differential effects on protected classes.[25] The requirements for the study went far beyond looking at inputs, and required the Fed and the FTC to study the outcomes of credit scoring on the availability and affordability of financial products.[26] The resulting study found that "the credit characteristics included in credit history scoring models do not serve as substitutes, or proxies, for race, ethnicity or sex," but that "[d]ifferent demographic groups have substantially different credit scores," and that "credit outcomes—including measures of loan performance, availability, and affordability—differ for different demographic groups."[27]

Disparate impacts cannot be avoided or disproven by dissecting a model feature-by-feature.[28] If defendants are able to block claims from proceeding based on this defense, plaintiffs will never have the chance to show that a model is discriminatory *despite* omitting close proxies for protected classes. Under HUD's existing disparate impact rule, when plaintiffs allege a prima facie disparity, both parties have the opportunity to engage in discovery and prove or disprove whether a model causes discriminatory outcomes and whether it is actually predictive of a legitimate and nondiscriminatory business interest. This fact-specific inquiry is necessary to root out algorithmic discrimination.[29] Without discovery, plaintiffs (and in some cases even defendants) will not have access to the artifacts required to assess discrimination, such as executable code, training data, system requirements documents, or validation test results.

[25] Fair & Accurate Credit Transactions Act of 2003, 117 Stat. 1952, Sec. 215 (2003), https://www.govinfo.gov/content/pkg/STATUTE-117/pdf/STATUTE-117-Pg1952.pdf.

[26] *Id.*

[27] Board of Governors of the Fed. Reserve Sys., Report to Congress on Credit Scoring & its Effects on the Availability & Affordability of Credit s–1 (Aug. 2007), https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/creditscore.pdf [hereinafter "FACT Act Report"].

[28] *See, e.g.*, Skanderson & Ritter, *supra* note 19.

[29] The need to test seemingly innocuous algorithms is demonstrated by a 2015 study that involved a machine learning algorithm designed to predict which hospital patients would develop pneumonia complications. Based on correlations in the training data, the algorithm detected a correlation between having asthma and having better survival rates from pneumonia. The researchers were puzzled by this correlation since asthma, as a respiratory condition, would seem likely to make outcomes worse for a patient who then contracted pneumonia. Ultimately, they concluded that patients with asthma were more likely to seek medical treatment early after detecting symptoms of pneumonia. Additionally, because hospitals typically send patients with asthma to intensive care, these patients were rarely logged as requiring further care in the hospital's records. These are the records with which the machine learning algorithm was trained, and as a result it produced flawed results. Thus, while people with asthma did in fact have statistically higher survival rates, the conclusion was that hospital personnel, on encountering a person with asthma and symptoms of pneumonia, should continue to treat that case as urgent rather than following the algorithms conclusion that this was a less urgent case. Rich Caruana et al., *Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission*, Proceedings of the 21st ACM SIGKDD Internat'l Conference on Knowledge Discovery & Data Mining, 1721–30 (2015), http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf.

B. There are no broadly applicable standards for determining whether an algorithmic model is "statistically sound" or "empirically derived," nor do these characteristics amount to nondiscrimination.

The third defense allows a defendant to show that a neutral third party has found that the model was "empirically derived" and "statistically sound" and that it "accurately predicts risk or other valid objectives."[30] This defense would allow actors to use discriminatory algorithms to make housing decisions as long as the models have predictive validity.[31] This would be a significant reversal of HUD's existing disparate impact rule. Under the existing rule, a defendant can engage in practices that cause disparate impacts *only if* the practice is necessary to achieve a legitimate business purpose, the business purpose itself is non-discriminatory, and there are no less discriminatory means of achieving it.[32] HUD proposes to circumvent this entire inquiry if a defendant can show that its model actually predicts a "valid objective" (HUD does not define valid objectives).

There are no broadly applicable standards for third parties or HUD to use to certify that an algorithmic model is "empirically derived" and "statistically sound," nor does HUD attempt to define these terms.[33] Algorithmic models designed to make complex predictions—such as who is likely to default on a loan, who is likely to click on an advertised offer, or who will be a reliable tenant—are inherently subjective. Basing such predictions in data does not make them automatically more correct or even more reliable. For example, social media websites typically use algorithmic models to prioritize and deliver personalized content to their users based on some measure of "relevance."[34] Since relevance is an abstract concept, it gets operationalized by measures such as a user's behavior (what they click on and

[30] 84 Fed. Reg. 42854, 42862.

[31] CFPB's Regulation B defines "empirically derived, demonstrably and statistically sound credit scoring system[s]" as those that are based on data about "creditworthy and non-creditworthy applicants," developed for the purpose of evaluating creditworthiness, and "validated using acceptable statistical principles and methodology." 12 C.F.R. § 1002.2(p). This standard only addresses statistical validity and says nothing about discrimination.

[32] 24 C.F.R. § 100.500.

[33] While Regulation B articulates a standard with respect to credit scoring, this standard would not readily map onto other types of models across industry sectors. *See* 12 C.F.R. § 1002.2(p). For example, the regulation relies on the explicit premise that creditors have a "legitimate business interest" in using historical data to assess creditworthiness for purposes such as "minimizing bad debt losses and operating expenses." *Id.* For more nascent practices, such as screening tenants based on social media and behavioral data or online advertising, neither HUD nor any third party has established guidance for what makes a model statistically sound or what kind of model use amounts to a legitimate business purpose. *See, e.g.*, TenantAssured, http://www.tenantassured.com/frequently-asked-questions/ (last visited Oct. 17, 2019). Even for credit scoring, adherence to Regulation B's definition alone would not mitigate disparate impacts. In fact, disparate impacts are incredibly likely when a model relies on comparing groups that have historically had access to economic opportunities or have been economically successful to those who have not and repeating those patterns. *See, e.g.*, Jeffrey Daster, *Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women*, Reuters (Oct. 9, 2018), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[34] *See, e.g.*, Dawei Yin et al., *Ranking Relevance in Yahoo Search*,
Proceedings of the 22nd ACM SIGKDD Internat'l Conference on Knowledge Discovery & Data Mining, 323–32 (2016), https://www.kdd.org/kdd2016/papers/files/adf0361-yinA.pdf.

interact with), characteristics of the content (e.g., whether it is news, when it was posted, and by whom), and other factors determined by the company (or learned by the algorithm from behavioral data).[35] Thus, there are thousands of ways to define and predict relevance. Yet these algorithms are likely to be analyzed under HUD's disparate impact rules. Advertising targeting practices have come under investigation by HUD and courts,[36] and one study found that advertisement delivery on Facebook can be skewed by race or gender even when the advertiser makes no decisions about how to target the ad.[37] How should a neutral third party determine whether an advertising algorithm is "statistically sound" in predicting which social media users should see a particular ad? HUD provides no guidance on how its proposed standards should be applied to real-world models impacting the availability and cost of housing.

A model can have predictive validity and cause the type of housing discrimination that HUD is tasked with preventing. Models may simply be reflecting disparities or patterns of discrimination that exist in the world. For example, traditional credit scores are widely used as predictors of credit risks, but the Fed and FTC's 2003 FACT Act report showed disparities in credit scores for protected classes.[38] Screening algorithms in housing often use criminal records as a factor in making risk-based predictions.[39] While data may reflect a correlation between criminal records and default risk, discriminatory practices in policing, housing, employment, and other sectors influence who will have a criminal record and what opportunities or obstacles they will face as a result. Any prediction that relies on criminal records is likely to disparately impact people of color, who are arrested at higher rates than whites.[40] HUD has acknowledged the risk of discrimination in relying on criminal records and has issued guidance stating that criminal record-based housing restrictions that cause disparate impacts are FHA violations:

> [C]riminal history-based restrictions on access to housing are likely to disproportionately burden African Americans and Hispanics . . . . [A] discriminatory effect resulting from a policy or practice that denies housing to anyone with a prior arrest or any kind of criminal conviction cannot be justified, and therefore such a practice would violate the Fair Housing Act.[41]

[35] *See, e.g.*, Adam Mosseri, Facebook Newsroom, Newsfeed Ranking in Three Minutes Flat (May 22, 2018), https://newsroom.fb.com/news/2018/05/inside-feed-news-feed-ranking/.

[36] *See* Complaint, *Dep't of Housing & Urban Development v. Facebook*, HUD ALG, FHEO No. 01-18-0323-8 (Mar. 28, 2019), https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf; Nat'l Fair Housing Alliance, Facebook Settlement, https://nationalfairhousing.org/facebook-settlement/ (last visited Oct. 17, 2019).

[37] Ali et al., *supra* note 14.

[38] FACT Act Report, *Supra* n. 27.

[39] *See, e.g.*, Conn. Fair Housing Ctr. v. CoreLogic Rental Property Solutions, D. Conn. 3:18-cv-00705-VLB (2019), http://www.ctfairhousing.com/PDFs/CoreLogicMTDOrder.pdf; TransUnion SmartMove, Criminal Report, https://www.mysmartmove.com/SmartMove/tenant-background-report.page (last visited Oct. 17, 2019).

[40] *See, e.g.*, Julia Angwin & Jeff Larson, *Bias in Criminal Risk Scores is Mathematically Inevitable, Researchers Say*, ProPublica (Dec. 30, 2016), https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say; https://hrdag.org/2016/10/10/predictive-policing-reinforces-police-bias/.

[41] Dep't of Housing & Urban Development, Office of Gen. Counsel Guidance on Application of Fair Housing Act Standards to the Use of Criminal Records by Providers of Housing and Real Estate-related Transactions 10 (Apr. 4, 2016), https://www.hud.gov/sites/documents/HUD_OGCGUIDAPPFHASTANDCR.PDF.

HUD's own interpretation of its existing rules makes it clear that statistical validity alone is not enough to overcome discrimination. Defendants whose uses of models cause disparate impacts should be subject to a merits-based inquiry into whether the models are the least discriminatory means of achieving a legitimate business purpose.

C. HUD's proposal to rely on "recognized third parties" and "intended uses" refers to standards that do not exist and would incentivize irresponsible practices.

The second defense HUD proposes would allow a defendant to show that "the challenged model is produced, maintained, or distributed by a recognized third party that determines industry standards, the inputs and methods within the model are not determined by the defendant, and the defendant is using the model as intended by the third party."[42] This defense ignores the reality of how algorithmic models are currently developed, maintained, and used across industries. There are no broadly applicable established standards for determining whether a model developer is a "recognized third party" or whether the model's users are using the model "as intended," and HUD provides no guidance on making these determinations. This defense would be particularly overbroad, since many companies do not develop their own models from scratch and rely on at least some third-party models. This proposed defense would simply shield most uses of algorithmic models from otherwise valid causes of action.

There are no broadly applicable established industry standards for developing and maintaining nondiscriminatory algorithmic models.[43] HUD does not specify which "industry" it is referring to, but models impacting housing opportunities could come from a range of industries, including the online advertising industry, social media and online search providers, the insurance industry, the banking industry, credit bureaus, data brokers, real estate services, rental and roommate matching services, and the short-term and vacation rental market. Even for the most sophisticated industry actors developing and providing access to models, processes for avoiding discrimination, documenting model and data specifications, and methods of constraining or overseeing third-party use remain experimental[44] and

[42] 84 Fed. Reg. 42854, 42862.

[43] Several methods for testing credit models for discrimination are in development and use by industry actors, but they are iterative and context-specific and ultimately focused on assessing the overall outcome of the system rather than invidual variables in isolation. *See* Skanderson & Ritter, *supra* note 19, at 15 ("[T]he focus of fair lending [in pre-screened marketing] is typically on the risk of disparate impact. The overarching objective of the fair lending analysis is to determine whether the impact of the *process as a whole* tends to exclude certain groups from credit offers disproportionately . . . .").

[44] *See, e.g.*, Sorelle A. Friedler et al., *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*, Proceedings of the Conference on Fairness, Accountability, and Transparency, 329–38 (2019), https://arxiv.org/abs/1802.04422; Skanderson & Ritter, *supra* note 19, at 34 ("No guidance has been published by the federal regulators of financial institutions regarding how the disparate impact risk of credit scoring systems should be tested or how large a disparate impact needs to be before it becomes a regulatory compliance concern.").

mostly non-public. For some complex models such as neural networks, the developer may not even be aware of all of the features and statistical relationships the model is relying on.

When one party develops a model and another party uses it, it is challenging for the developer to ensure that the model will be used only for appropriate decisions in accordance with its design and capabilities. Even if downstream users do not determine the inputs used to develop the model or its underlying logic, they will necessarily introduce new information to the model that will influence its outputs, possibly undermining assumptions made during development. For example, face recognition is currently being used to control access to some apartment buildings.[45] Studies using face recognition and analysis systems as intended have demonstrated race and gender disparities resulting from the fact that the faces on which the systems were used represented a different race and gender distribution than the faces on which the systems were trained.[46] In other cases, models trained on data from one locality have produced discriminatory outcomes when used in other localities with different demographics.[47]

HUD's existing disparate impact rule incentivizes companies to carefully evaluate their own models as well as third-party models they use, which in turn incentivizes third-party developers to provide documentation. The NPRM would reverse this incentive by allowing actors that use third-party algorithms to put their heads in the sand or even to use known discriminatory models with impunity.

II. The proposed rule would undermine the purpose of disparate impact and create regulatory confusion.

The Fair Housing Act obligates HUD to end housing discrimination and "affirmatively further fair housing."[48] In its 2013 disparate impact rule, HUD acknowledged that recognizing disparate-impact liability allows the agency to accomplish both of these goals by enforcing against policies that result in segregated housing but where no evidence of intentional discrimination can be shown.[49] The NPRM would undermine HUD's statutory obligation by making it nearly impossible for plaintiffs to allege a prima facie case of disparate impact and disincentivizing fair housing practices.

[45] *See* Elizabeth Kim, *Brooklyn Landlord Wants to Install Facial Recognition Tech at Rent-Stabilized Complex*, Gothamist (Mar. 25, 2019), https://gothamist.com/news/brooklyn-landlord-wants-to-install-facial-recognition-tech-at-rent-stabilized-complex.

[46] *See, e.g.*, Joy Buolamwini & Timnit Gebru, Gender Shades, http://gendershades.org/ (last visited Oct. 17, 2019); Natasha Singer, *Amazon is Pushing Facial Technology that a Study Says Could be Biased*, N.Y. Times (Jan. 24, 2019), https://www.nytimes.com/2019/01/24/technology/amazon-facial-technology-study.html; Congressional Black Caucus, Letter to Amazon About Facial Recognition Technology (May 24, 2018), https://cbc.house.gov/news/documentsingle.aspx?DocumentID=896; Jacob Snow, Am. Civ. Liberties Union, *Amazon's Face Recognition Falsely Matched 28 Members of Congress with Mugshots* (July 26, 2018), https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[47] *See, e.g.*, Phillip Knox & Peter Keifer, Nat'l Ctr. for State Courts, The Risks and Rewards of Risk Assessments, https://www.ncsc.org/microsites/trends/home/Monthly-Trends-Articles/2017/The-Risks-and-Rewards-of-Risk-Assessments.aspx (last visited Oct. 17, 2018).

[48] 42 U.S.C § 3608(d); 78 Fed. Reg. 11460, 11465, 11477 (2013).

[49] 24 C.F.R. § 100; 78 Fed. Reg. 11460, 11465, 11477 (2013).

Algorithmic models raise precisely the risk of discrimination that disparate impact is supposed to address. As this comment has discussed, models can reflect and perpetuate discriminatory patterns despite using no obvious proxies for protected classes or intentionally discriminatory design. If these biases are not mitigated, models can be used to make systematically discriminatory decisions at scale.

Plaintiffs whose discrimination claims challenge decisions that are based on algorithmic models will be unable to meet the new prima facie pleading burdens HUD proposes. Algorithmic models are notoriously difficult to examine from the outside. Their training data, design, logic, and outputs are typically kept secret. Some models do not expose their logic of operation even to their developers. HUD's existing burden-shifting framework would allow parties to engage in discovery once a prima facie disparity is alleged, allowing more of the models' inner workings, use cases, and outcomes to come to light. Under the NPRM, the parties would never get to this point. The algorithmic defenses would prevent almost every claim from proceeding to litigation. Thus, the NPRM is an enormous departure from HUD and federal court precedent.

The rule would also create unnecessary regulatory confusion for industry. The existing disparate impact rule provides clarity for companies about the nondiscrimination standards to which their models will be held. The test is based on whether the model results in discriminatory outcomes. The NPRM would create vague new rules, without any guidance or industry standards to rely upon, requiring companies to identify which inputs are and are not "close proxies" for protected classes and which parties are "recognized third parties." The "close proxy" standard could actually inhibit attempts to affirmatively further fair housing through initiatives to ensure that housing opportunities are available on an equitable basis and that affordable housing is not segregated into certain neighborhoods.[50]

Conclusion

The NPRM is an unprecedented departure from decades of HUD and federal court precedent enforcing against disparate-impact housing discrimination. The defenses HUD proposes for algorithmic models have no basis in law or in data or computer science. For the foregoing reasons, we oppose the NPRM and request that HUD continue to analyze algorithmic disparate impacts under its existing burden-shifting framework.

[50] Protected class information can be critical for developing fair models. *See, e.g.*, Dino Pedreschi, Salvatore Ruggieri & Franco Turini, *Discrimination-Aware Data Mining*,
Proceedings of the 14th ACM SIGKDD internat'l conference on Knowledge Discovery & data mining, 560–68 (2008), http://pages.di.unipi.it/ruggieri/Papers/kdd2008.pdf; Zachary Lipton, Alexandra Chouldechova & Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, 32nd Conference on Neural Information Processing Systems (2018), https://arxiv.org/pdf/1711.07076.pdf; Betsy Anne Williams, Catherine F. Brooks & Yotam Shmargad, *How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications*, 8 J. Info. Policy 78 (2018), https://www.jstor.org/stable/pdf/10.5325/jinfopoli.8.2018.0078.pdf; Skanderson & Ritter, *supra* note 19.

Respectfully submitted,

Center for Democracy & Technology

Center on Privacy and Technology at Georgetown Law

Color of Change

Data & Society

Electronic Frontier Foundation

Free Press Action

Media Mobilizing Project

National Hispanic Media Coalition

New America's Open Technology Institute

Open MIC (Open Media and Information Companies Initiative)

Public Knowledge

United Church of Christ, OC Inc.

Upturn

Samuel T. Brandao, Clinical Instructor, Tulane Civil Rights and Federal Practice Clinic*

Robin Burke, Chair, Department of Information Science, University of Colorado, Boulder *
Member, GRAIL (Governance Research in Artificial Intelligence Leadership) Network*

Christina Drummond, Program Manager and Senior Analyst, Program on Data and Governance, Moritz College of Law*

Kelly Capatosto, Senior Data and Policy Specialist, Kirwan Institute for the Study of Race and Ethnicity*

Dennis Hirsch, Professor of Law and Faculty Director of the Program on Data Governance, The Ohio State University Mortiz College of Law*

Sarah Igo, Andrew Jackson Professor of History, Vanderbilt University*

Krisitan Lum, Member, GRAIL (Governance Research in Artificial Intelligence Leadership) Network*

Shira Mitchell, Member, GRAIL (Governance Research in Artificial Intelligence Leadership) Network*

Shobita Parthasarathy, Professor of Public Policy & Women’s Studies, University of Michigan*
Director, Science, Technology, and Public Policy Program, University of Michigan*

Suresh Venkatasubramanian, Professor, School of Computing, University of Utah*
Member of the Board, ACLU of Utah*
Member, GRAIL (Governance Research in Artificial Intelligence Leadership) Network*

**Institutional affiliation is provided for identification purposes only and does not constitute institutional endorsement.*