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: Comment on Proposed Rule: HUD's Implementation of the Fair Housing Act's Disparate Impact Standard; Docket No. FR-6111-P-02

I. INTRODUCTION

This comment focuses on two issues for which HUD specifically seeks input in its Notice of Proposed Rulemaking (NPRM): (1) whether the NPRM's disparate impact test aligns with the ruling in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc.*, 135 S.Ct. 2507 (2015) (*Inclusive Communities*); and (2) whether the safe harbor defenses for defendants who rely upon algorithmic models are proper. The answer to both questions is no. The NPRM does not bring HUD's disparate impact rule into alignment with *Inclusive Communities*. Rather, it distorts the substance of the traditional three-part disparate impact test that the Court recognized, as well as the long-standing assignment of burdens of proof to plaintiffs and defendants. In so doing, the NPRM weakens the Fair Housing Act and undermines Congress's intent "to eradicate discriminatory practices" in housing. *Id.* at 2521. Meanwhile, the algorithm defense is wholly new, harmful, unnecessary, and based on clear misunderstandings about how algorithmic decision-making works. It also has no justification under any law, and should be eliminated entirely.

We write as law professors and attorneys with expertise in housing law and algorithmic accountability. Andrew Selbst is a Postdoctoral Scholar at the Data & Society Research Institute, and an incoming Assistant Professor of Law at UCLA School of Law. Drawing on his background in both engineering and law, he has written several articles on various aspects of algorithmic accountability, and disparate impact in particular,[1] and regularly provides expert advice to legislative bodies on these issues. Michele Gilman is the Venable Professor of Law at the University of Baltimore School of Law, where she teaches Administrative Law and directs the Civil Advocacy Clinic, which has a robust housing

[1] *See* Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, Proceedings of FAT* '19, the ACM Conference on Fairness Accountability and Transparency 59 (2019); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham Law Review 1085 (2018); Andrew D. Selbst & Julia Powles, *Meaningful Information and the Right to Explanation*, 7 Int'l Data Privacy L. 233 (2017); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 Ga. L. Rev. 109 (2017); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 677-94 (2016).

law practice representing tenants. She is currently a faculty fellow at the Data & Society Research Institute, researching the impact of algorithmic decision-making on low-income communities, a topic on which she regularly writes.[2]

II. THE NPRM DISTORTS THE SUPREME COURT'S DISPARATE IMPACT DOCTRINE AND WILL UNDERMINE THE REMEDIAL PURPOSE OF THE FAIR HOUSING ACT

A. The Three-Part Disparate Impact Test Explained

In *Inclusive Communities*, the Supreme Court upheld the theory of disparate impact liability in cases brought under the Fair Housing Act. Disparate impact cases challenge practices that have a "'disproportionately adverse effect on minorities.'" 135 S. Ct. at 2513. Unlike disparate treatment cases, disparate impact cases do not require proof that "'the defendant had a discriminatory intent or motive.'" *Id*. at 2513. Disparate impact claims are central to eliminating "zoning laws and other housing restrictions" that keep minorities out of certain neighborhoods "without any sufficient justification." *Id*. at 2521-22. Disparate impact claims can also "counteract unconscious prejudices and disguised animus" that perpetuate segregated housing patterns. *Id*. at 2522.

Inclusive Communities affirmed the longstanding, three-part burden-shifting framework for adjudicating disparate impact claims, imported from Title VII jurisprudence. This three-part framework was set forth by HUD in its existing disparate impact rule, issued in 2013.[3] At that time, HUD made clear that it was "not establishing new substantive law [but] rather, this final rule embodies law that has been in place for almost four decades." 78 Fed. Reg. 11460, 11462. In *Inclusive Communities*, the Court quoted from HUD's existing rule at length without suggesting that it conflicted with the holding in the case or needed revamping. 135 S. Ct. at 2514-15. To the contrary, the Court referred to the three steps of the burden-shifting framework throughout the opinion. It also affirmed the judgment of the Court of Appeals for the Fifth Circuit's decision in the case below; the Fifth Circuit explicitly applied HUD's 2013 three-part burden-shifting test. 747 F.2d 275 (2014). Moreover, as the Court made clear, the three-part test is the same standard that has been used since 1971 in the employment discrimination context and that is directly analogous to housing cases. 135 S. Ct. at 2516.[4]

In *Inclusive Communities*, the Court explains the three-part test for adjudicating a fair housing disparate impact claim as follows:

---

[2] T*he Surveillance Gap: The Harms of Extreme Privacy and Data Marginalization*, 42 NYU Rev. of L. & Soc. Change 253 (2019) (with Rebecca Green); *Privacy, Poverty and Big Data: A Matrix of Vulnerabilities for Poor Americans*, 95 Wash. U. L. Rev. 53 (2017) (with Mary Madden, Karen Levy & Alice Marwick); *The Class Differential in Privacy Law*, 77 Brooklyn L. Rev. 1389 (2012).

[3] Implementation of the Fair Housing Act's Discriminatory Effects Standard, 78 Fed. Reg. 11460 (2013).

[4] "To be sure, the Title VII framework may not transfer exactly to the fair-housing context, but the comparison suffices for present purposes." 135 S. Ct. at 2523.

Step 1: To establish a prima facie case, the plaintiff must allege facts or produce statistical evidence that a defendant's specific practice caused or will cause a discriminatory effect. *Id*. at 2514, 2523.

Step 2: The burden then shifts to the defendant, who must prove "the valid interest served by [its] policies," *id*. at 2522, which is done by establishing "that the challenged practice is necessary to achieve one or more substantial, legitimate, nondiscriminatory interests." *Id*. at 2515. (This step is "analogous to the business necessity standard under Title VII and provides a defense against disparate-impact liability." *Id*. at 2522).

Step 3: The burden then shifts back to the plaintiff, who must prove that there is "'an available alternative…practice that has less disparate impact and serves the [entity's] legitimate needs.'" *Id*. at 2518.

B. The "Artificial, Arbitrary, and Unnecessary" Standard is Not Supported by the Fair Housing Act or *Inclusive Communities*.

Under the "new burden-shifting framework" in the NPRM, Plaintiffs will have to plead that the practice or policy they are challenging is "arbitrary, artificial, and unnecessary to achieve a valid interest or legitimate objective." This is a complete rewriting of disparate impact law.

The disparate impact doctrine was first recognized by the Supreme Court in *Griggs v. Duke Power Co*., 401 U.S. 424 (1971). In *Griggs*, the Court explained that the purpose of Title VII was "the removal of artificial, arbitrary, and unnecessary barriers to employment when the barriers operate invidiously to discriminate on the basis of racial or other impermissible classification." *Id*. at 431. This is a statement of the *Griggs* Court's reading of Congressional intent. As the Court explained, Congress sought to remove all of these barriers to opportunity, not just the intentionally imposed one. In *Inclusive Communities*, the Court states three times that disparate impact liability is designed to remove "artificial, arbitrary, and unnecessary" barriers. 135 S. Ct. at 2522, 2524. Each time this phrase appears, it is a direct quote from *Griggs. Inclusive Communities* was literally and figuratively echoing the *Griggs* Court: it recognized that the purpose of the FHA was to remove these barriers. When it comes to the substantive standard for identifying and removing these barriers, the Court in *Inclusive Communities* imported it directly from current disparate impact law, which stems from a different part of the *Griggs* decision and has since evolved. The plaintiff's burden in making a prima facie case of disparate impact has always been to show that "a challenged practice caused or predictably will cause a discriminatory effect," *Id*. at 2514, and it is then the Defendant's burden to prove the "valid interest served by their policies." *Id*. at 2522.

The NPRM treats these quotations from *Griggs* as demanding a new substantive standard, but that is simply incorrect. The *Griggs* court established early on that the "touchstone" of disparate impact was "business necessity," 401 U.S. at 431, a defense often-litigated and clarified in the years since. The *Griggs* court offered many different possible standards for the new business necessity test:

> A challenged employment practice must be "shown to be related to job performance," have a "manifest relationship to the employment in question," be "demonstrably a reasonable measure of job performance," bear some "relationship to job-performance ability," and/or "must measure the person for the job and not the person in the abstract."[5]

Notably, none of the *Griggs* substantive standards is so narrow that it only covers practices shown to be "artificial, arbitrary, or unnecessary," and in the decades since 1971, neither the Supreme Court nor the Equal Employment Opportunity Commission has ever treated the business necessity standard this way. And accordingly, the *Inclusive Communities* Court does not treat the FHA's "valid interest" standard this way either.

The NPRM's discussion of a new "robust causality" standard similarly takes descriptive language from *Inclusive Communities* and twists it. The Court wrote that

> a disparate-impact claim that relies on a statistical disparity must fail if the plaintiff cannot point to a defendant's policy or policies causing that disparity. A robust causality requirement ensures that '[r]acial imbalance ... does not, without more, establish a prima facie case of disparate impact" and thus protects defendants from being held liable for racial disparities they did not create.

Inclusive Communities, 135 S. Ct. at 2523 (quoting Wards Cove Packing Co. v. Atonio, 490 U.S. 642, 653 (1989)). But that describes precisely the existing test. The plaintiff's prima facie case requires statistical evidence of discriminatory effects, but the second and third steps satisfy the "without more" limitation. It is almost tautological that if a defendant's policies do not cause the disparate impact, there should be no liability. But the existing test is designed such that the defendant may argue that a statistical disparity is inadequate in his defense. As with the rest of the case, the *Inclusive Communities* Court is merely importing to the Fair Housing Act what is clearly spelled out in Title VII.[6] Rather than recognize this, the NPRM treats robust causality as an excuse to impose new burdens on plaintiffs. Robust causality already exists; to use the phrase to make plaintiffs' lives harder is contrary to existing law.

---

[5] Linda Lye, Comment, *Title VII's Tangled Tale: The Erosion and Confusion of Disparate Impact and the Business Necessity Defens*e, 19 Berkeley J. Emp. & Lab. L. 315, 321 (1998) (footnotes omitted) (quoting *Griggs*, 401 U.S. at 431–36).

[6] *Cf.* 42 U.S.C. 2000e-2 (k)(ii) ("If the respondent demonstrates that a specific employment practice does not cause the disparate impact, the respondent shall not be required to demonstrate that such practice is required by business necessity.")

C. The NPRM Rewrites the Supreme Court's Holding by Reassigning the Burden of Proving a "Valid Interest Served" from the Defendant to the Plaintiff.

The NPRM audaciously claims to write a "new burden-shifting framework." 84 Fed. Reg. 42858. The first thing to note is that HUD does not have the power to rewrite Supreme Court precedent or the Fair Housing Act. An agency's decision to contradict a Supreme Court case that interprets an unambiguous statute is contrary to law and will not warrant *Chevron* deference[7] -- or any level of deference whatsoever.

The burdens in disparate impact are well established. The plaintiff's burden in making a prima facie case has always been to show that "a challenged practice caused or predictably will cause a discriminatory effect." 135 S. Ct. at 2514. Once a plaintiff meets this burden, it is then the Defendant's burden to prove the "valid interest served by their policies." *Id.* at 2522. Nowhere does the Court hold that the burden of *disproving* a defendant's valid interest belongs to the plaintiff's prima facie case. This is nothing new, as HUD's 2013 rule requires defendants to prove in Step Two that "the challenged practice is necessary to achieve one or more substantial, legitimate, nondiscriminatory interests." *Id.* at 2515.

The Court's placement of the burden on defendants to establish the validity of their policies and practices makes perfect sense. Plaintiffs cannot be expected to prove a negative, ie., that there is no rebuttal to their prima facie case. Plaintiffs cannot be expected to assume a new pleading and proof burden that has no analogy or precedent in any other area of non-discrimination law. Plaintiffs cannot be expected to explain the internal thinking or reasoning of governmental bodies or private developers – especially at the Complaint drafting stage before there has been the fact-finding provided through discovery.

In the NPRM, HUD acknowledges, as it must, that plaintiffs "will not always know what legitimate objective the defendants will assert in response to the plaintiff's claim." In such cases, the pleading requirement is lower. But, according to the NPRM, where the "policy or practice has a facially legitimate objective," the plaintiff must allege that it is "artificial, arbitrary, and unnecessary." How is a plaintiff to judge whether the defendant's intention is clear on its face? The NPRM does not say, thus opening the door to costly litigation about the clarity of defendant's expression in adopting a new policy or practice. It makes no sense as a matter of cost or efficiency to place a burden of proof on the plaintiff that is entirely within the control of the defendant.

---

[7] *Chevron U.S.A. Inc. v. Natural Resource Defense Council, Inc.*, 467 U.S. 837 (1984) (holding that where there is a statutory ambiguity left by Congress, courts must defer to an agency's reasonable interpretation of a statute it administers). There is no ambiguity for HUD to interpret in this case because (1) the Supreme Court has ruled on the issue, and (2) Congress "accepted and ratified" the unanimous holdings of the Courts of Appeals finding disparate impact liability – and doing so under the three-part burden shifting framework – in the many cases decided before Congress amended the Fair Housing Act in 1988. 135 S.Ct. at 2520.

In *Inclusive Communities*, the Court reasoned that when Congress amended the Fair Housing Act in 1988,[8] it was aware that all nine Courts of Appeals to have addressed the issue had recognized disparate impact claims under the FHA. By amending the Act without altering settled precedent, "Congress accepted and ratified the unanimous holdings of the Courts of Appeals finding disparate-impact liability." *Id.* at 2520. As the Court explains, this is a standard method of statutory interpretation. By the same reasoning, Congress also approved in 1988 the three-part burden shifting framework that courts have used for over forty years to assess disparate impact liability. Because the three-part test is part of the statute, HUD does not have the authority to rewrite it.

In fact, looking at the subsequent history of Title VII, it is hard to imagine a clearer rejection of this new burden allocation. On June 5, 1989, the Supreme Court decided *Wards Cove Packing Co. v. Atonio*, 490 U.S. 642 (1989). *Wards Cove* was the latest in a series of cases that weakened the business necessity defense. Most importantly for present purposes, the *Wards Cove* Court reallocated the burden of proof in the business necessity defense to the plaintiff, just as this NPRM seeks to do. *Id.* at 659. Until then, disparate impact was a judicial interpretation of Title VII, and had not been codified in statute. But this new burden allocation was the last straw for Congress. In 1991, Congress finally codified disparate impact in the Civil Rights Act of 1991, Pub. Law 102-166, 105 Stat. 1071, *codified at* 42 USC § 2000e–2. While Congress said almost nothing about the specific business necessity standard, the one thing that it made crystal clear was that the entire purpose was to overrule *Wards Cove* and revert the state of disparate impact to "the law as it existed on June 4, 1989." 42 USC § 2000e–2 (k)(1)(C).[9] Thus, the "new" burdens that this NPRM seeks to implement were in fact the very impetus for the codification of a disparate impact standard that roundly rejected them. In *Inclusive Communities*, the Court held that the disparate impact test in the Fair Housing Act is the very same that exists in employment law, the same one that Congress codified in 1991.

By imposing a "new" burden on plaintiffs to prove that defendants' policies or practices are not artificial, arbitrary, or unnecessary, the NPRM rewrites the settled three-part burden-shifting framework that governs disparate impact claims in discrimination lawsuits. This "new" burden risks undermining the purpose of the FHA to end segregation in housing because it will make it difficult, if not impossible, for plaintiffs to meet their burden of proof. Simply put, plaintiffs are not mind readers and thus cannot anticipate and rebut all the possible justifications a defendant may put forward to support a challenged

---

[8] In the amendments, Congress banned familial discrimination and created three exemptions from FHA liability.

[9] The statute quotes that directly with respect to the third step of the disparate impact test, but legislative history is clear with respect to the entire test. The bill itself states as follows: "No statements other than the interpretive memorandum appearing at Vol. 137 Congressional Record S15276 (daily ed. Oct. 25, 1991) shall be considered legislative history of, or relied upon in any way as legislative history in construing or applying, any provision of this Act that relates to Wards Cove." Pub. Law 102-166, 105 Stat 1071. The interpretive memorandum, in turn states: "The terms 'business necessity' and 'job related' are intended to reflect the concepts enunciated by the Supreme Court in *Griggs v. Duke Power Co.*, and in the other Supreme Court decisions prior to *Wards Cove Packing Co. v. Atonio*." 137 Cong. Rec. § 15,276 (Oct. 25, 1991) (interpretive memorandum).

policy or practice. HUD claims that the purpose of the NPRM is to "bring HUD's disparate impact rule into closer alignment with the analysis and guidance provided in *Inclusive Communities* …". 84 Fed. Reg. 42857. At the same time, HUD admits it is writing a "new" burden-shifting framework. Both these propositions cannot be true. The truth is that the NPRM's definition of plaintiffs' prima facie burden is new. Because it conflicts with the very Supreme Court case that HUD claims to be interpreting, it should not be adopted.

D. The NPRM Gives Defendants a Free Pass to Discriminate by Eliminating Defendants' Burden of Proof and Creating Unwarranted Safe Harbors for Defendants.

In the NPRM, HUD proposes to alter the longstanding, respective proof burdens of plaintiffs and defendants in FHA disparate impact cases. As discussed above, the NPRM adds to plaintiff's burden by requiring plaintiff to prove that the challenged practice is "arbitrary, artificial, and unnecessary"; and as discussed in this section, the NPRM simultaneously extinguishes the defendant's burden of proof. The only conclusion that can be drawn from this revision is that HUD is seeking to protect defendants in FHA cases, despite the Court's reaffirmance of the importance of the FHA and its observation that "'no dire consequences'" have resulted from several decades of disparate impact cases. *Id.* at 2525.

The NPRM destroys the careful balance the three-step, burden-shifting framework, proposing to lower the defendant's Step Two burden from a burden of proof to a burden of production. Under the NPRM, the defendant no longer has to prove a legitimate reason for its challenged practice – it simply has to state one. 84 Fed. Reg. 42863, proposed 100.500(d)(1)(ii) (the defendant's burden is "producing evidence showing that the challenged practice or policy advances a valid interest…"). This is in direct conflict with the Court's discussion of Step Two in *Inclusive Communities*. As the Court states, "housing authorities and private developers [must] be allowed to maintain a policy **if they can prove** it is necessary to achieve a valid interest." *Id.* at 2523 (emphasis added).

Although the NPRM's discussion of the respective proof burdens is confusing and unduly complicated (thus ensuring years of wasteful litigation), it is clear that the NPRM's change will make it harder for plaintiffs to establish disparate impact, because plaintiffs would have the burden of proving not only their affirmative case, but also to prove a negative, ie., that no defense is possible, while giving defendants the power to defeat a case with only minimal evidence.

E. The NPRM Creates Safe Harbors for Defendants Despite a Lack of Congressional or Judicial Authority.

The NPRM creates two "safe harbors" for defendants, allowing them a complete defense if their discretion is limited by a third party (such as a federal, state, or local law; or a binding adjudicative or administrative requirement) or if they relied on an algorithmic model. Proposed 100.500(d)(2). These safe harbors are completely unnecessary because these are factors that defendants have long been able to raise at Step Two, in proving the business necessity for their challenged practice. Creating safe harbors contradicts HUD's 2013 disparate impact rule and years of court rulings that stress the

uniqueness of each disparate impact case. As the 2013 rule states, the Step Two inquiry is "case-specific [and] fact-based." 78 Fed. Reg. at 11470; *see also id*. at 11471 (defenses are "fact-specific" and "must be determined on a case-by-case basis"). In the 2013 rule, HUD rejected the idea of creating safe harbors, including "examples of tenant screening criteria such as rental history, credit checks, income verification, and court records that would be presumed to qualify as legally sufficient justifications." *Id.*

In the 1998 FHA Amendments, Congress created three exemptions from disparate impact liability: (1) real estate appraisers can consider factors other than the protected categories; (2) landlords can exclude people with criminal drug convictions; and (3) landlords can restrict the number of occupants in a dwelling. These safe harbors constrain disparate impact liability in narrow circumstances. If Congress wanted to create additional safe harbors, it could have done so. It did not, even though computer modelling was in effect in 1988 as well as discretionary limitations facing government agencies and private entities in the housing markets. If Congress wanted HUD to create additional safe harbors, it could have directed HUD to do so. It did not. Instead, Congress preserved the three-step burden-shifting framework for all cases outside the statutory safe harbors. If the Supreme Court believed that the FHA mandated additional complete defenses in the disparate impact context, it could have said so in *Inclusive Communities*. It did not. Instead, the Court acknowledged the three-step test as the governing paradigm.

III. THE NRPM'S TREATMENT OF ALGORITHMS IS UNINFORMED, UNNECESSARY, AND HARMFUL, EFFECTIVELY CREATING AN IMMUNITY FOR DEFENDANTS WHO USE ALGORITHMS.

The NPRM offers three safe-harbor defenses for algorithm users. If the defendant:

> (i) Provides the material factors that make up the inputs used in the challenged model and shows 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;
>
> (ii) Shows 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) Shows 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;

then the plaintiff fails to make a prima facie case.[10]

These defenses exhibit a total lack of understanding about the nature of algorithmic discrimination and the state of the industry. Permitting a defense showing that the inputs are not "substitutes or close proxies for protected classes" and that the model is predictive misses the point that the algorithms can cause discrimination *despite those facts*. Permitting a defense showing that the model is the fault of some third party that determines industry standards creates functional immunity because there is no mention that the industry standard be anti-discriminatory, and the third party likely cannot be sued under the Fair Housing Act. Permitting "an objective and unbiased neutral third party" to validate the truth of the first defense and that the algorithm is "statistically sound" fails to account for algorithmic discrimination for the same reason the first defense does.

Though the NPRM claims that the "section is not intended to provide a special exemption for parties who use algorithmic models," 84 Fed. Reg. 42859, that is precisely the effect of the rule as written. Moreover, there is no need for these new defenses, because they can be raised under the existing affirmative defense (Step Two) within disparate impact doctrine. Ironically, then, these defenses are "artificial, arbitrary, and unnecessary barriers" to plaintiffs' ability to enforce disparate impact law. They should be eliminated entirely, and converted to guidance about how the traditional business necessity defense applies to a world in which algorithmic decision-making occurs.

A. Algorithms Are Not Neutral and Discrimination That Results from an Algorithm Is Caused by the Algorithm Design.

As the NPRM explains, the idea behind these algorithmic defenses is that "a successful defense … would demonstrate the lack of a robust causal link between the defendant's use of the model and the alleged disparate impact." 84 Fed. Reg. 42859. Ignoring all the methodological problems discussed below, this is simply incorrect. To understand why, it is necessary to understand how algorithmic modeling based on machine learning works in general. What follows is a highly simplified primer.

The goal of these models is to predict some unobservable trait, like likelihood of defaulting on a future loan. The trait is unobservable either because it is not directly measurable or because it is in the future; in the case of most issues with housing, like ability to pay rent, it is the latter. The way that algorithmic modeling does this is to look for patterns in existing data. With enough data about "good" and "bad" housing candidates, a landlord can look for similar traits and—the theory goes—select for the good ones. The "good" and "bad" candidates are examples of what are called class labels. This also requires information about their features. These could be, for example, income, prior defaults, how much they previously paid in rent, or how long they were at past addresses. (Note that we are not endorsing these as permissible features to examine for discrimination purposes, as some may be

[10] These defenses appear in a new Section detailing when plaintiffs fail to make out a prima facie case, 84 Fed. Reg. 42862, but they operate as affirmative defenses for which plaintiff bears the burden. *Id.* at 42863. This makes little sense. If the defenses exist at all, they should be treated as affirmative defenses.

proxies for protected class, but they are examples of features that might be available and relevant to housing.). The computer would then see which features are held in common by the people who are labeled as "good" candidates, and build the model, which essentially weighs the importance of each of those features. Then, when making predictions, the model will be fed a new candidate with known features, but an unknown class label. The model will compare the candidate's features to the known "good" outcomes and will assign some probability of being a "good" or "bad" candidate in the end.

The reason that statistical validation of a model does not negate causation is that the creation of these algorithmic models requires many subjective decisions. The first, and perhaps most important, is the selection of what to optimize for. A computer cannot actually understand "good" or "bad" candidate, so a subjective choice must be made about how to make it computable. The variable a computer optimizes for is called the "target variable." For just one of infinitely many possible examples, a landlord might want candidates who are 70% likely not to miss a rent payment in two years. This choice is itself highly subjective: A mortgage lender could choose to optimize for a percentage that will make loans available to the greatest number of people, but might also optimize for the greatest profit. If the algorithm takes into account default rates, it might recommend giving out unaffordable loans with high interest to maximize default, if that maximizes profit. This would be a very different kind of result.

These subjective choices can lead to discrimination. In 2012, the Wall Street Journal reported about an employment screening company that developed an algorithm that discovered that distance from work was predictive of job tenure; the closer a worker lives, the more likely they are to stay with the company. However, the company eliminated this variable from the algorithm because employees' distance from work can be racially disparate due to segregated housing patterns in America.[11] In this case, distance from work is probably not a substitute or close proxy to race, yet the model relied heavily on it, with discriminatory results. The key point here is that the discrimination was the result of the subjective software designer's choice to optimize for job tenure (a decision that was ultimately scrapped because of its disparate impact). By contrast, if the algorithm had instead optimized for customer satisfaction ratings (one employer operated a call center), then distance to work likely would not have mattered as much, and the result might have been less racially disparate.

Once a target variable is chosen, there are many other subjective choices to be made. One example is what data to collect and analyze. A common practice is to use a "convenience sample," made of whatever data the algorithm designer has lying around. This sample will not be representative, and will have errors that are often worse along lines of protected class. There are also human decisions that go into data cleaning, fixing gaps and errors, or deciding when data is simply too unreliable to use. Sometimes data is hard to get or expensive, but will hugely improve the model's accuracy and/or

---

[11] *See* Joseph Walker, *Meet the New Boss: Big Data*, Wall St. J. (Sept. 20, 2012), http://www.wsj.com/news/articles/SB10000872396390443890304578006252019616768, *see also* Dustin Volz and National Journal, *Silicon Valley Thinks It Has the Answer to Its Diversity Problem*, The Atlantic, Sept. 26, 2014, https://www.theatlantic.com/politics/archive/2014/09/silicon-valley-thinks-it-has-the-answer-to-its-diversity-problem/431334/.

equity. Where better quality data is extremely costly to obtain, this might be a defensible choice, but a choice it is nonetheless. These subjective decisions are absolutely necessary to model building and are unavoidable.[12]

While many of the choices will be defensible, defensible choices are no less subjective, and the degree of discrimination would be a direct result of the choice of data that is used. Under traditional disparate impact doctrine, a defendant would argue that the cost makes the version of the algorithm trained on a convenience sample a business necessity, while a plaintiff would be able to argue that a less discriminatory alternative—an algorithm based on better data—was available. Thus, this is a merits question about whether the choices amount to illegal discrimination.

Ultimately, the entire purpose of using algorithms is to find patterns in data to make predictions about the future. There is therefore no static ground truth to compare against, and no basis to suggest that the algorithm is discovering some unknown objective truth. If it discriminates, the discrimination is caused by the choice of mechanism used, just as with any traditional tool commonly considered in disparate impact cases. The NPRM's assumption that algorithms break the causal link between a user's decision to adopt an algorithm and any resulting disparate impact is incorrect.

B. Whether Inputs to Algorithmic Models Are "Substitutes or Close Proxies for Protected Class" Is Not an Appropriate Test for Discrimination.

The discriminatory effect of algorithms cannot be determined by reference solely to their inputs. While disparate impact doctrine is certainly implicated where the inputs of algorithmic models are "substitutes or close proxies for protected classes," that is neither unique to algorithms nor the primary concern with them. Algorithms find patterns in data that people could not otherwise easily identify and sort candidates based on those patterns. Where individual inputs like zip code might be close proxies for protected classes, algorithms will often find patterns based on interactions *between* variables. Take the hiring example above: while a person's distance from work has some set correlation with race, the importance of that variable to the eventual model depends on all the other variables in the model and what is being optimized for. Thus, the model is based on all the variables interacting, and the specific output may or may not end up depending on any one input variable.

Over the last several years, computer science researchers have developed different methods to detect the influence of specific features in a model to the output, after the model is built.[13] This is a

---

[12] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Cal. L. Rev. 671, 677-94 (2016).

[13] *See e.g.* Philip Adler et al., *Auditing Black-Box Models for Indirect Influence*, 54 Knowledge and Info. Systems 95 (2018); David Baehrens et al., *How to Explain Individual Classification Decisions*, 11 J. Machine Learning Res. 1803 (2010); Anupam Datta et al., *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems*, *in* Proceedings of The 2016 Ieee Symposium On Security & Privacy 598 (2016); Andreas Henelius et al., *A Peek into the Black Box: Exploring Classifiers by Randomization*, 28 Data Mining & Knowledge Discovery 1503 (2014); Marco Tulio Ribeiro et al., *"Why Should I Trust You?" Explaining the Predictions of Any Classifier*, *in*

difficult problem precisely because the influence of particular input factors to the end result is not obvious *a priori*. It is therefore true both that a model can include close proxies for protected class and not have discriminatory results and that a model can be discriminatory without any known proxies for protected class.

This is why the discourse around algorithmic discrimination has so often turned to a need for explainable algorithms.[14] The discrimination that results from the choice of algorithm will often be difficult or impossible to understand without understanding the specific choices made within algorithm design, and without the ability to manipulate those choices, testing for more or less discrimination. Simply looking at inputs is not enough to ferret out discrimination.

Beyond the general inappropriateness of looking solely at inputs, the rule as written has several related deficiencies. What counts as a substitute or close proxy is not defined in the NPRM. How close a proxy must it be? Would zip code count? Does it have to be a well-known proxy? What if one input is discovered to be a proxy during the process of algorithm design? How is context taken into account? If a particular community is more geographically segregated than another community, then zip code will be a closer proxy for race in the former context; is an input proxy-ness fact-dependent? According to the NPRM, whether something is a proxy should be determined separately from its predictive power, but sometimes a single variable might be both predictive and a proxy: how are these treated? In the traditional disparate impact test, the use of such variables would lead a plaintiff to claim that there are less discriminatory alternatives available, but the rule as written does not permit such a claim.

C. An Algorithm's Statistical Validity Does Not Ensure That It Is Not Discriminatory.

The concerns with algorithmic discrimination arise precisely because one can have both apparently statistical validity and discriminatory results. Statistical validity of an algorithmic model is usually assessed by randomly subdividing the training dataset into training data and test data; the test data is often called "holdout data."[15] This method will demonstrate statistical validity in the sense that the model will predict the results of the holdout data with fidelity. But if the dataset is not representative of the community over which decisions are being made, then the holdout data will be just as skewed as the training data, as it is just a partition of the training data. Thus the model will appear statistically valid, yet still be discriminatory.

For another example, return once again to the hiring example above. The algorithm could have been optimized for job tenure or for something else, like customer feedback ratings. In either case, a well-designed model would be statistically valid. Yet, as described above, one would have a more

---

Proceedings of The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135 (2016).

[14] *See, e.g.*, Margot E. Kaminski, *The Right to Explanation, Explained*, 34 Berkeley Tech. L.J. 189, 191 n.3, 192 n.8 (collecting sources).

[15] *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653, 684–88 (2017).

discriminatory result than the other. This is because statistical validity does not mean that a model is not discriminatory.

D. The Third-Party Defense is Contrary to Law.

There is no basis under the law for a defendant to be able to disclaim liability because they purchase an algorithmic tool from a third party. It is simply immaterial that the defendant would not have intended the discrimination, as the very essence of a disparate impact claim is that the effect, not the intent, is discriminatory. Whether they create the discriminatory algorithm or purchase the discriminatory algorithm from a third party, the defendant has "otherwise ma[d]e unavailable or den[ied]" housing in exactly the same way. 42 U.S.C. § 3604(a). It may be that a court would eventually find joint liability on behalf of the housing defendant and a software developer, but that is a matter between the algorithm designer and the purchaser; the Fair Housing Act and *Inclusive Communities* are clear and unambiguous that disparate impact liability applies to the defendant with no safe harbor here.

E. There Is No Such Thing as a "Recognized Third Party That Determines Industry Standards."

Even were the third-party defense not wholly contrary to established law, the construction of it here makes little sense for several reasons. First, there is no existing set of industry standards around algorithmic decision-making. While several industry groups such as the IEEE and Partnership on AI are attempting to come up with standards for some parts of the algorithmic process, such as documentation,[16] to the best of our knowledge, there is no entity currently contemplating generalized industry standards with the intent of certifying algorithms.

Second, even were generalized standards to develop, the body that determines industry standards would very likely not be the vendor "produc[ing], maintain[ing], or distribut[ing]" algorithmic products. Standard-setting bodies are usually not companies that are selling the products that must comply to the standards, if for no other reason than the inherent conflict of interest that would make it difficult for them to set standards for themselves and their competitors.

Third, the NPRM does not specify the industry to which these standards apply. Context is everything. If standards emerge, they are likely to be different in housing, in credit, in banking, in employment, and in policing. Is the industry standard something that combines all of them under one "algorithm" umbrella, or are they particular to housing and lending, etc.? Or are they subdivided further, perhaps into big and small banks? Or in-house models versus commercial off-the-shelf systems (COTS)? The NPRM does not answer these questions.

F. The Existence of Industry Standards Does Not Imply an Absence of Discrimination.

The NRPM does not provide any sense of what these industry standards might be. Most importantly, the NPRM does not specifically envision that they are standards that contain anti-bias or

[16] *See e.g.*, ABOUT ML: Annotation and Benchmarking on Understanding and Transparency of Machine Learning Lifecycles, https://www.partnershiponai.org/about-ml/

anti-discrimination requirements. While some industry actors are interested in developing standards that are indeed non-discriminatory, there are plenty of companies who would simply develop whatever will make them the most money, with the least risk of legal liability. This NPRM, as written, would thus increase the chance that an industry standard will develop that cares nothing about discrimination, because it signals that liability will be hard to come by. Given that this rule is ostensibly about implementing disparate impact, permitting a defense that refers to a (non-existent) industry standard that is not about preventing discrimination is particularly perverse.

Even assuming the industry standard refers to some certification of anti-discrimination, when one digs further into the importance of context, it becomes apparent that the very idea of an industry standard COTS is out of touch with the reality of algorithmic systems. Algorithms are trained on a specific limited dataset. That data is usually comprised of people with different demographics. When a so-called "fair-ML" algorithm is optimized for non-discrimination on some population, it will have to be retrained on a different dataset when used in another location. Said differently, a landlord in New York City purchasing a certified non-discriminatory algorithm would not be purchasing the same algorithmic system as a landlord in Louisville or Santa Fe. The differing demographics would render the fairness guarantees invalid without retraining, so either it cannot meet the fairness standard or it is not a COTS.[17]

G. The Independent-Audit Defense Suffers from the Same Flaws as the First Defense.

The last defense essentially repeats the requirements of the first defense, but allows a defendant to show that those requirements have been satisfied by audit, rather than demonstrating them themselves. Showing compliance by audit is not inherently problematic, but the underlying requirements mirror the first defense, and thus have the flaws described above in sections III (a)–(c).

H. If the Intent Were Truly to Adopt the Existing Disparate Impact Standard, Rather than Unlawfully Alter It, These Additional Defenses Are Unnecessary.

Though the NPRM claims it is only bringing HUD policy into line with disparate impact doctrine, these affirmative defenses clearly depart from it. As described above, the Fair Housing Act contains safe harbors, so Congress knows how to write them in when it wants to. But there really is no need for additional safe harbors. The business necessity defense in traditional disparate impact doctrine is somewhat flexible. Landlords and banks are justifying algorithm use by saying it is the most efficient and most effective way to ensure that they distribute housing or loans to the best applicants. Under the traditional business necessity defense, defendants could produce evidence to support precisely this claim. For example, they could submit an audit of the algorithm demonstrating that other versions of the algorithm did not perform as well or were prohibitively expensive. The third prong of the test would be satisfied by similar evidence: when plaintiffs argue that there is a less discriminatory alternative

[17] *See* Andrew D. Selbst, et al., *Fairness and Abstraction in Sociotechnical Systems*, Proceedings of FAT* '19: the ACM Conference on Fairness, Accountability, and Transparency, 59, 61.

available, evidence of testing several versions of the algorithm, as well as non-algorithmic solutions, could satisfy these defense claims.

That is, just as this NPRM envisions that audits can help defendants fend off discrimination claims, so can they under traditional disparate impact law. The only difference is that the audits have to show things relevant to disparate impact - they must satisfy the substantive requirements of the business necessity and alternative practice prongs of the test. So either this rule is substantively changing the disparate impact standard—which HUD lacks the statutory authority to do—or these algorithmic safe harbors are unnecessary because they simply require the same evidence that would satisfy the existing second and third prongs of the disparate impact test. Therefore, either way, the safe harbors should be eliminated.

IV. CONCLUSION

In sum, despite the NPRM's claim, it does not align with the Supreme Court's decision in *Inclusive Communities*. The Court sought to preserve disparate impact theory and its substantive standards and methods for proof. The NPRM rewrites the law in an area where there is no ambiguity. The overarching goal of the NPRM appears to be to limit the availability of fair housing remedies by increasing plaintiffs' burdens while giving defendant's multiple avenues for avoiding liability. Further, the NPRM misunderstands how algorithms operate and thus risks immunizing a vast swath of decision-making with discriminatory effects.

Although the legal and algorithmic arguments made in this Comment can seem abstract, it is essential to remember that housing policies and practices have real-life consequences that impact countless people. Where people live impacts where they go to school, where they can work, their safety and security, and their physical and mental health. Despite the Fair Housing Act's announcement of "a clear national policy against discrimination in housing," discrimination and segregation in housing continues. Disparate impact cases are essential to changing this dynamic.

Significantly, disparate impact cases have led to successful outcomes in cases challenging mortgage providers that offered subprime mortgage loans to Hispanic and African-American borrowers, while offering prime loans to similarly situated white borrowers; insurance companies that refused to sell insurance to apartment owners who rent to tenants who participate in federally subsidized voucher programs; cities that blocked the construction of affordable housing developments; cities that concentrated affordable housing in minority neighborhoods; city zoning laws that required single-family lots, thereby limiting multifamily housing; public housing authorities that used local residency preferences in their subsidized housing programs, thereby excluding minorities; and more. These and other disparate impact cases make a difference in people's lives and opportunities. For example, numerous studies[18] establish that reducing segregated housing patterns is associated with

[18] A survey of relevant studies is in Jonathan Zasloff, The Price of Equality: Fair Housing, Land Use, and Disparate Impact, 48 Columbia Human Rights L. Rev. 98, 104-08 (2017).

better health for African-Americans, higher rates of intergenerational mobility,[19] and reductions in the gap between test scores of white and black students.[20] For all these reasons, HUD should continue to enforce its 2013 disparate impact rule to ensure progress in fair housing.

Respectfully submitted,

Andrew Selbst
Postdoctoral Scholar, Data & Society Research Institute
36 W. 20th St., 11th Floor
New York, NY, 10011
andrew@datasociety.net

Michele Gilman
Venable Professor of Law, University of Baltimore School of Law
Faculty Fellow, Data & Society Research Institute
University of Baltimore School of Law
1420 N. Charles Street, Baltimore, MD 21201
michele@datasociety.net

[19] Raj Chetty et al., Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States, 129 Q.J. Econ. 1553, 1608-11 (2014).
[20] David Card & Jesse Rothstein, Racial Segregation and the Black- White Test Score Gap, 91 J. Pub. Econ. 2158 (2007).