Office of the General Counsel
Rules Docket Clerk
Department of Housing and Urban Development
451 7th Street, SW, Room 10276
Washington, DC 20410-0001.

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

We are a group of computer scientists, social scientists and legal scholars who are writing to express our concern about the proposed amendments to HUD's implementation of the Fair Housing Act, and in particular those amendments related to the use of algorithms.

We believe that the proposed amendments are based on a failure to recognize how modern algorithms can result in disparate impact, even in the absence of discriminatory intent, and how subtle the process of auditing algorithms for bias can be. In fact, the proposed amendments only explicitly address algorithms in prescribing defenses for their use; it does not address their potential harms or unintended consequences. Moreover, these amendments would allow lenders and other defendants to avoid responsibility to the point that disparate impact liability would effectively disappear wherever algorithms are used. As a result, the proposed amendments would move HUD farther from, not closer to, the Supreme Court upholding disparate impact liability in *Texas Department of Housing and Community Affairs v. Inclusive Communities Project, Inc*. (*Inclusive Communities*).

We have four arguments we wish to make.

1. **To ensure that an algorithm does not have disparate impact, it is not enough to show that individual input factors are not "substitutes or close proxies" for protected characteristics as in Paragraph 100.500(c)(2)(i).**

We appreciate HUD's recognition of this issue. Zip codes were used in the racially discriminatory practice of redlining. Similarly, proxies—that is, variables that are strongly correlated with protected characteristics such as race—can result in policies that, while they appear neutral on their face, have significant disparate impact.

However, it is insufficient to test whether individual factors are proxies. The nature of algorithms is to find subtle patterns and correlations in large data sets, including identifying complex relationships between factors. As a result, disparate

impact can occur if *any combination of input factors, combined in any way,* can act as a proxy for race or another protected characteristic.[1] An algorithm's output is a holistic combination of all of its inputs. It is simply not possible to, as the proposal suggests, "break down the model piece-by-piece and demonstrate how each factor considered could not be the cause of the disparate impact." Suggesting this as a standard reflects a fundamental misunderstanding about how algorithms work.

It could easily be the case that each individual input factor passes a statistical test based on, say, statistical significance or the four-fifths rule, but that their cumulative effect when combined creates significant bias due to their interactions with each other. Examining crime data, for instance, computer scientists have found[2] that "in practice multiple variables combine to result in a stronger proxy than any of the individual variables." This is true even for simple types of algorithms like those based on linear or logistic regression, and even more so for more complicated algorithms that combine factors in complex and nonlinear ways.[3]

Of course, a defendant such as a landlord, broker, or lender could protest that even though some set of factors can be combined to create bias, their algorithm does not combine these factors in this or a similar way, and therefore that their algorithm as a whole is unbiased. This brings us to our next point.

2. **It is impossible to audit an algorithm for bias without an adequate level of transparency or access to the algorithm.**

If an algorithm is transparent—that is, if we understand what factors it uses, and how these factors are weighted and combined—then it can be convincingly audited for disparate impact. But if the defendant using paragraph 100.500(c)(2)(i) refuses to disclose what the algorithm does with its input factors—what calculations it performs on them to produce its recommendation or risk score—this is far more problematic and violates the standard rule of model replication across both the social and hard sciences.

There are auditing methods for "black box" algorithms, namely those where we can see the inputs and output of an algorithm without knowing its inner workings. But these methods require us to have, at a minimum, "black box access"

---

[1] See e.g. Feldman et al., "Certifying and removing disparate impact." Proc. 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015

[2] Yeom, Datta, and Fredrikson, "Hunting for discriminatory proxies in linear regression models." Proc. 32nd International Conference on Neural Information Processing Systems, 2018.

[3] Indeed, more complex algorithms can detect more complicated relationships between factors, making them capable both of higher accuracy and of more subtle forms of discrimination against protected classes.

to the algorithm: that is, to be able to probe its behavior by giving it various inputs and seeing how it responds.

For instance, suppose we wish to know whether a risk assessment algorithm for lending is influenced by a particular factor in a way that leads to disparate impact. We could test this by asking the algorithm about fictional loan applicants where this factor has been adjusted to reduce the difference between its distribution in a protected class and in the population as a whole.[4]

However, algorithm providers are typically not willing to have independent analysts or plaintiffs probe their algorithm with such inputs. In the private sector, terms of service often prevent this type of access explicitly, regarding it as "tinkering" that could allow an algorithm to be reverse engineered. It seems unlikely that the defendant would provide even black-box access to their algorithm unless compelled to do so.

Moreover, the designers of proprietary algorithms typically claim that their inner workings are trade secrets protected under intellectual property law. A plaintiff could ask that the source code of an algorithm, or black-box access to it, be provided as part of the discovery process, perhaps under a protective order or non-disclosure agreement. But judges have tended to defer to trade secret protections, even in criminal cases when arguments for greater transparency have been made under *Frye* or *Daubert*.[5] In any case, unless the plaintiff satisfies the requirements for pleading a prima facie case, they will not even proceed to discovery. In many, if not most cases, plaintiffs will need the information a defendant may provide under 100.500(c)(2)(i) at the time of filing the complaint.

Another mechanism by which algorithms can have disparate impact is if the data they use contains inaccuracies, and if those inaccuracies go unchecked—especially given that inaccuracies such as criminal records are not evenly distributed across the population. Therefore, in addition to transparency about how algorithms treat data across statistical groups, we also need transparency at the individual level. In analogy with the Fair Credit Reporting Act, applicants must be allowed access to the data about them that is used to approve or deny their application, and have the ability to meaningfully contest, update, or refute that data if it is inaccurate. Moreover, applicants should be entitled to at least a partial

---

[4] Adler et al., "Auditing Black-box Models for Indirect Influence." *Knowledge and Information Systems* 54(1), 95–122 (2018), and Marx et al., "Disentangling influence: using disentangled representations to audit model predictions." Preprint, https://arxiv.org/abs/1906.08652. In addition to testing whether an algorithm "rel[ies] in any material part on factors that are substitutes or close proxies for protected classes," in some cases this black-box access also makes it possible to construct algorithms that are just as accurate but where disparate impact has been reduced or eliminated, addressing the plaintiff's burden under paragraph (d)(1)(ii).

[5] Roth, "Machine testimony." *Yale Law Journal* 126:1972 (2017).

explanation if they are denied, i.e., the most important factors that led to the algorithm's decision. Algorithms that opaque, proprietary, or not amenable to this type of transparency and explainability should come under special scrutiny for potential bias.

To sum up, a lack of transparency or access to an algorithm and the data it uses would make it difficult or impossible for a plaintiff to establish a "robust causal link," thus undermining one of the main tests laid out in *Inclusive Communities*. This objection holds even more strongly for third-party algorithms, bringing us to our third point.

3. **Allowing defendants to deflect responsibility to proprietary third-party algorithms effectively destroys disparate impact liability.**

Paragraph 100.500(c)(2)(ii) would allow the defendant to avoid all liability by using a model "produced, maintained, or distributed by a recognized third party…" While the proposal claims that this "is not intended to provide a special exemption for parties who use algorithmic models," it would do exactly that. Such algorithms are already commonly in use, and will become more so. We expect lenders, buyers, brokers, and renters of property to use this defense as a matter of course, making disparate impact claims against a wide variety of defendants impossible to prove. Indeed, this section would strongly incentivize potential defendants to use third-party algorithms to avoid liability.

In addition, we typically do not even know what input factors a third-party algorithm uses. The third party could easily include data gathered through contracts with additional companies such as consumer preferences, social media, the credit ratings of others in the applicant's social network, and so on. It should go without saying that these additional factors can and do act as substitutes for protected characteristics.

In lieu of challenging the defendant, the proposal suggests that "suing the party that is actually responsible for the creation and design of the model would remove the disparate impact from the industry as a whole." At present this suggestion is impractical. We know of very little case law that would hold the designer of an algorithm liable for disparate impact caused by use of their algorithm. After all, third-party algorithm providers can argue that they do not sell or rent properties, or grant or deny loans or mortgages; they merely provide a software tool that makes recommendations to those who do.[6]

[6] One of the only cases along these lines is *Connecticut Fair Housing Center v. CoreLogic Rental Property Solutions*, 369 F.Supp.3d 362 (D. Conn. 2019), but it focuses largely on the disparate impact of using criminal records as part of a background check.

Note also that the accuracy of an algorithm and the extent to which it discriminates both depend on the context in which it is applied, including the demographic properties of each population. Even if an algorithm is "validated" on a benchmark data set from one place and time, it may have disparate impact when applied to another. An algorithm which meets statistical definitions of fairness in the Midwest, for instance, might have disparate impact when applied in the South, or vice versa. In the criminal justice arena, this has led to calls for "local re-validation" of algorithms wherever they are in use.[7] Thus a one-time validation based on some yet-to-be-defined "industry standard" does not suffice. Algorithms must be continually re-validated and re-audited for bias as populations vary between regions and change over time.

Finally, the language of paragraph 100.500(c)(2)(iii), releasing the defendant from liability if the algorithm "has been subjected to critical review and has been validated by an objective and unbiased neutral third party," invites a number of questions. What measures of accuracy and disparate impact will the third party use in their analysis? Will their validation appear in the peer-reviewed scientific literature? Will it be based on publicly available benchmark data sets? Can it be independently verified and reproduced by other scientists? Without clear answers to these questions, we believe that algorithms can and will lead to increased disparate impact.

4. **The proposed regulation fails to take into account the cumulative impact of multiple users of algorithms that result in disparate impact on protected classes where no individual user has liability under the proposed regulation.**

We are concerned that the proposed regulation does not appreciate the cumulative effect of disparate impact caused by the use of algorithms. The purpose of the FHA was to eliminate racial discrimination and segregation in housing in the United States. The Supreme Court in *Inclusive Communities* stated that "Much progress remains to be made in our Nation's continuing struggle against racial isolation." "This Court acknowledges the Fair Housing Act's continuing role in moving the Nation toward a more integrated society." *Inclusive Communities* was decided five years ago and these statements remain true today. The proposed regulation is so focused on assuring that mortgage lenders and landlords can make profits, it loses sight of the potential for algorithms to rapidly reverse that progress. Mortgage lending and renting properties are legitimate businesses but those engaged in them are not permitted to discriminate even if doing so would be profitable.

[7] see e.g. National Center for State Courts, *The Risks and Rewards of Risk Assessments.*

The Court in *Inclusive Communities* upheld disparate impact claims but required safeguards against the imposition of liability where a governmental agency or private party did not cause the disparate impact, particularly in the case of one-time decisions. The decision did not address algorithms which by their nature are used in numerous decisions and transactions. Not only may the use of algorithms in mortgage lending and renting properties cause disparate impact but such use may perpetuate a more segregated society in various ways. First, if multiple lenders and landlords use an algorithm that adversely affects a protected class and no one user is liable for this disparate impact, the magnitude of the disparate impact is multiplied. Second, the nature of such algorithms is to create stereotypes rather than individualized determinations. It is not about "you," it is about the statistical groups into which the algorithms place you. Algorithms are trained to find patterns in historical data and establish statistical classes of individuals in terms of creditworthiness and good tenants. Those classes are not static. A class may be expanded as new information is available to the algorithm. A person who defaults on a loan or makes a late payment on a loan or rent not only impacts the class into which he or she is grouped but also the classes in which people associated with them are grouped, i.e. neighbors, social media contacts, relatives, co-workers, etc. Thus, stereotyping by algorithm may lead to more racial segregation contravening the very purpose of the FHA.

There is another concern. The presence of significant cumulative disparate impact indicates that a mortgage lending or housing market is concentrated or less than competitive. Antitrust law is of limited usefulness in combating the disparate impact resulting from firms using algorithms since the plaintiff must prove collusion among competing firms or that a firm is a monopolist. HUD is, however, in a position to regulate the cumulative disparate impact, especially when algorithms facilitate collusion.[8] The Court in *Inclusive Communities* acknowledged that the purpose of the FHA is to address the aggregate societal impact of disparate impact. The proposed amendments will provide no barrier to the cumulative disparate impact resulting from stereotyping by algorithm.

To conclude: we are entering an algorithmic age, in which many decisions to sell, rent, approve loans, or otherwise make housing available will be made based on recommendations generated by automated methods. While these methods are potentially more accurate and objective than human decision-making, they also have the potential to affect some groups adversely, especially when they are based on historical data and thus perpetuate historical patterns. Whether or not their

[8] Ariel Ezrachi and Maurice E. Stucke, Artificial Intelligence & Collusion: When Computers Inhibit Competition, 2017 U. Ill. L. Rev. 1775.

designers or users have discriminatory intent, algorithms, by definition, have no intentions at all. This makes it difficult to prove disparate treatment, even though algorithms clearly have the capacity to discriminate. Our best recourse is to vigorously subject them to the test of disparate impact.[9] These proposed amendments, if adopted, would make that far more difficult, and would move HUD away from the Supreme Court's decision in *Inclusive Communities*.

Sincerely,
Interdisciplinary Working Group on Algorithmic Justice

[with affiliation given for identification only]

Cristopher Moore, Professor, Santa Fe Institute
Alfred Mathewson, Professor and former Dean, University of New Mexico School of Law
Elizabeth Bradley, Professor, Computer Science Department, University of Colorado, Boulder, and the Santa Fe Institute
G. Matthew Fricke, Research Assistant Professor, University of New Mexico Computer Science Department and Center for Advanced Research Computing
Mirta Galesic, Professor, Santa Fe Institute
Joshua Garland, Postdoctoral Fellow, Santa Fe Institute
Melanie Moses, Professor, Computer Science Department, University of New Mexico
Kathy Powers, Associate Professor, Department of Political Science, Senior Fellow, Center for Social Policy, University of New Mexico
Sonia M. Gipson Rankin, Professor, University of New Mexico School of Law
Gabriel R Sanchez, Professor, Department of Political Science, Director, Center for Social Policy, University of New Mexico

[9] Barocas and Selbst, "Big Data's Disparate Impact." *California Law Review* 104:671 (2016).