AI and Housing Discrimination: The Case of Mortgage Applications

Abstract

Issues surrounding bias and discrimination in housing markets have been acknowledged and discussed both in the literature and in practice. In this study, we investigate this issue specifically in the context of mortgage applications through the lens of an AI-based decision support system. Using the data provided as a part of the Home Mortgage Disclosure Act (HMDA), we first show that ethnicity bias does indeed exist in historical mortgage application approvals, where black applicants are more likely to be declined a mortgage compared with white applicants whose circumstances are otherwise similar. More interestingly, this bias is amplified when an off-the-shelf machine-learning model is used to recommend an approval/denial decision. Finally, when fair machine-learning algorithms are adopted to alleviate such biases, we find that the "fairness" actually leaves all stakeholders – black applicants, white applicants, and mortgage lenders – worse off. Our findings caution against the use of machine-learning models without human involvement when the decision has significant implications for the prediction subjects.

Keywords: Artificial Intelligence, Bias, Causal Inference, Discrimination, Mortgage

1 INTRODUCTION

Housing discrimination, an issue where prejudicial treatments affect the minority's ability to rent or buy a house, has been a significant societal issue in the past several decades [1]. In the United States, it has been estimated that housing discrimination costs African American $5.7 billion and Hispanics $3.4 billion every three years, resulting in significant economic disadvantage to minorities [2]. Housing discrimination manifests in several different ways. One common way, which is the focus of this paper, is discrimination in the mortgage market [3, 4]. Specifically, this issue is observed as racial gaps in loan denials and mortgage costs among mortgage applicants. This issue has attracted interest from researchers and practitioners over the past decade because even though the housing discrimination issue appears to be in decline overall, discrimination in the mortgage market is remarkably persistent, contributing to housing inequality and racial disparities among minorities [4].

In this study, we investigate how the use of artificial intelligence (AI) in the mortgage application process affects the issues surrounding bias and discrimination in the mortgage market. Our study is inspired by the increasing adoption of AI-based decision support systems among decision-makers in various contexts, including the justice system [5], credit card fraud detection [6], and job application screening [7]. As such, it is important to study the implications of the adoption of an AI-based decision support system in the mortgage application process, which historically suffers from racial discrimination, from the value/risk proposition perspective. In that regard, we empirically examine how a common off-the-shelf AI model (i.e., an AI model that is publicly and immediately available to use) contributes to the issue of bias and discrimination in the

mortgage application process. In addition, we study the implications of the adoption of a fair AI model (i.e., an AI model that are designed to provide fairness) in this context. Specifically, since the literature has shown that fair AI models are mostly designed to provide fairness in prediction results while the behavior of prediction subjects could be asymmetrically affected [8], we study how the adoption of a fair AI model in the mortgage application process impacts the welfare of stakeholders. Our paper uses historical mortgage application data that was disclosed as a part of the Home Mortgage Disclosure Act (HMDA). This legislation requires financial institutions that offer mortgage-related products to disclose anonymous information regarding mortgage applications and corresponding final decisions. The data contain applicants' mortgage types, income levels, loan values, ethnicities, races, gender, and other personal and related geographical statistics. We treat this dataset as a training dataset and develop a supervised learning-based AI model that relies upon an off-the-shelf machine learning algorithm to classify whether each mortgage application is approved or declined. With the developed model, we first demonstrate that when a common AI model is used, unsurprisingly, it inherits the ethnicity bias that exists in the training data, producing results that are biased against black applicants. Here, we employ multiple methods to ensure that the bias detected is not merely a correlation between the ethnicity and the approval decision, but rather, the causal inferences can be established (i.e., we can causally attribute the discrepancy in the approval decision to the ethnicity of the applicants). In this exercise, we also find that the AI model not only inherits bias from the training dataset but actually amplifies that bias. Subsequently, we adopt a popular fair AI model that was designed to mitigate AI prediction bias. We find that when a fair AI model is used to make approval decisions, fairness is improved as intended. However, the cost of such fairness is substantial, and it is paid by the mortgage lender.

Our findings are significant for future research, as well as carrying real-world implications. First, the direct use of a common AI model to approve mortgage applications without human intervention should be discouraged since it amplifies the bias from the historical training dataset. Second, when the fair AI model is used to mitigate such a bias, it is the mortgage lender who pays for the cost of fairness. These key findings caution against the use of AI models, even fair ones, to automatically process mortgage applications. In the next section, we survey prior literature in the areas of housing discrimination, AI bias, and bias detection and mitigation in AI.

2 RELATED LITERATURE

2.1 Bias and Discrimination in AI Predictions

The increased use and development of AI models that leverage big data has raised legal concerns regarding the implementation of discriminatory analytics systems. For instance, issues of algorithmic biases have emerged around applications of AI, including in mortgage application approval. Prior works have discussed how latent bias in training data induces discrimination and affects eventual predictions. For example, Favaretto et al. discuss the use of big data, unfair discrimination, and inequality and disparity that appear in credit scoring processes [9]. In the same way, Schneider discusses how Latinos are affected by unfair algorithms that are used to select tenants, which ultimately contribute to housing discrimination [10]. While there is a lot of empirical evidence of bias, discrimination, and inequality predictions from AI models, there also exist studies that show the opposite. For example, big data and AI can help reduce human bias when they are properly implemented [11], such as the use of AI as an inconsistency detector [12]. As such, it is

critically important not only to detect biases, but also to uncover the underlying mechanisms that drive them and identify potential mitigation solutions. In that regard, Podsakoff et al. summarize common bias sources and the effects of human behaviours from a psychological perspective [13]. The theories and methods proposed in this work have been widely used in other studies to prove the capacity of big data and the influence of big data in decision-making with respect to human bias [14]. Nevertheless, one of the concerns that frequently appear in studies that identify bias in big data is whether the bias detected between a sensitive variable and an outcome variable is a mere correlation, or truly causal.

With the need to ensure that causal inferences can be established for bias detected in data ecosystems, several recent works have proposed methods that satisfy such requirements. Two methods that are commonly used in the literature are called Fair on Average Causal Effect (FACE), and Fair on Average Causal Effect on the Treated (FACT) [15]. Both methods are developed to establish causal inferences regarding fairness in a prediction process using the Rubin-Neyman potential outcomes framework [16] where FACE focuses on fairness in a population while FACT focuses on fairness in a protected group. Both FACE and FACT operate by building logistic regressions and weighting sensitive variables such as gender, ethnicity, or race, and then comparing the outcome values with the sensitive variables [8]. For example, FACE can be used to examine gender bias by comparing the outcome variable (e.g., salary) between female and male employees. Meanwhile, FACT has an advanced step to find the best matches of the observations with different sensitive variables. For instance, in the same task described earlier, FACT would attempt to discover the matching pairs of female and male employees who have similar observable characteristics that may affect the outcome variable (e.g., education background and experience), and compares their salary levels to ascertain whether gender bias exists. FACE and FACT methods are demonstrated to be reliable based on multiple public datasets, including the adult income dataset from the UCI repository, NYC Stop and Frick data, and COMPAS dataset with recidivism prediction [17]. At the same time, another study proposes an algorithm that leverages the random forest algorithm to estimate the Causal Average Treatment Effect (CATE), called Causal Forest [18].

Apart from methods that detect and identify bias with causal inferences, there also exist other bias detection techniques that are based on a different school of thought. For instance, there exist techniques that identify bias based on the disparities in prediction accuracy between prediction subjects with different sensitive variables [19, 20]. As an illustration, in the prediction where the target variable is salary, these techniques argue that if the algorithm predicts the salary of male employees more accurately than the salary of female employees, then the gender bias exists in the prediction task. These techniques are available in several publicly available packages, including Microsoft Fairlearn [19] and IBM Fairness 360 Suite [21].

2.2 Housing Discrimination

In the context of housing and especially mortgage applications, past research has shed light on discriminatory biases. However, discriminatory lending practices and financial inequality are not new problems [22]. Mortgage discrimination and inequality are historical problems that were brought to the fore by several lawyers and economists in the context of the 2008 global financial crisis [23, 24]. Some have even argued that racial inequality in financial and housing markets constitutes one of the background conditions that contributed to the exponential growth of the subprime loans, which eventually led to the financial crisis of 2008 [25]. Evidence shows that during the pre-subprime era, those with low incomes, especially blacks

and Hispanics that are not qualified for traditional mortgages, were obliged to pay their own way. They were charged high prices for financial services. Banks have often offered services to nonstandard customers indirectly through the provision to firms operating subprime services. While restrictions on price competition provided banks with competitive incentives to maximize their customer bases, racial and other forms of discrimination still led them to ignore and underserve some areas. Many lower income households, and especially minority households, were subject to financial exclusion because they did not have access to the same financial services that the average-income household could obtain. Often, non-bank suppliers of financial services offer high-price substitutes for services that traditional banks provide at a lower cost. These gaps reflect long-standing patterns of racial and ethnic inequality and discrimination.

While many lenders viewed these terms as justified because they considered minorities customers high-risk, several economists and lawyers have argued that these methods opened the way for an exploitative financial model in which subprime loans became a common predatory lending practice [23]. Originally, subprime mortgages were offered in minority neighborhoods whose residents had been denied access to more traditional mortgages. By 2000, subprime loans grew by 900 percent. A study of HDMA data of the year 2000 found that both Hispanics and African Americans were twice likely as whites to receive subprime loans (even if they qualified for traditional mortgages) [26]. While this is beyond the scope of this paper, it is worth noting that the exponential growth of unconventional financing and the resulting financial crisis went hand in hand with financial and labor de-regulation [24]. In fact, the regulatory attitude of the regulators, including the Federal Reserve and Congress, was made clear through their passivity, as evidenced in the pre-subprime era [23]. Recent evidence has shown that while housing discrimination is in decline, inequality in financing persists [4]. As discussed in this study, minorities are more likely to be denied mortgages. Thus, automating such processes reinforces unfair lending practices and results in unethical and illegal discrimination. Accordingly, it may be argued this there is a strong case to be made for specific regulation on the algorithmic bias.

3 DATA AND RESEARCH METHODOLOGIES

In this section, we first describe the dataset that serves as the primary source of our empirical study. Subsequently, we discuss the research methodologies involved.

3.1 Data

The analyses in this study are based on the open dataset from the official website of the Federal Financial Institutions Examination Council (FFIEC). Particularly, we use the dataset that is released as the part of the Home Mortgage Disclosure Act (HMDA) in 2019. HDMA requires financial institutions to release, report, and disclose the information related to mortgages to the public. The data are made available while applicants' and users' privacy are protected. The dataset includes variables that are related to mortgage applications such as loan type, loan purpose, and loan terms. In addition, the dataset also includes demographic information of mortgage applicants such as applicants' race, ethnicity, and gender. In this study, in accordance with prior literature, we treat the variable race as the sensitive variable.

In the dataset, the variable race consists of five categories: (1) American Indian or Native, (2) Asian, (3) Black or African American, (4) Native Hawaiian or Other Pacific Islander, and (5) White. We select Black or African American as the focal group and merge other races as well as "information not provided" and "not

applicable" as the opposite group. In addition, the final decision of the mortgage application is set as a binary outcome variable, which is either approved or declined. After standard pre-processing processes, the final dataset consists of a binary sensitive variable of the race (i.e., Black or others), a binary outcome variable (approved or declined), and 25 other variables. In the 25 selected predictors, there are 7 numerical variables and 18 categorical variables. The summary statistics of the numerical variables are reported in Table 1. In total, the dataset consists of 2,861,627 observations.

Table 1: Summary Statistics of Numerical Variables

Variables	Column description	Mean	Median	Minimum	Maximum	Standard deviation
loan_amount	Amount of the loan or the amount (in dollars) applied for	214,742.12	155,000	5,000	64,000,000	252,658.67
loan_term	Number of months after which the legal obligation will mature or terminate	317.35	360	1	3,600	78.86
property_value	Value of the property (in dollars) relied on that secures the loan	400,226.62	295,000	5,000	64,000,000	503,178.95
tract_minority_population_percent	Percentage of minority population to total population for tract, rounded to two decimal places	32.75	24.52	0	100	26.91
tract_to_msa_income_percentage	Percentage of tract median family income compared to MSA/MD median family income	112.05	108.00	0	507	47.24
tract_owner_occupied_units	Number of dwellings, including individual condominiums, that are lived in by the owner	1,410.16	1,286	0	19,536	908.52
tract_median_age_of_housing_units	Tract median age of homes	35.26	34.00	0	76	18.70

3.2 Research Methodologies

3.2.1 Detecting Algorithmic Bias with Causal Inferences

In our study, we leverage multiple methods to detect algorithmic bias in the AI prediction process as well as establish causal inferences in the detection process. All methods we employ have a specific requirement that

the structure of the dataset contains three components: one binary sensitive variable, one outcome variable, and other predictor variables. The sensitive variable constitutes the treatment in the study (i.e., the value separates the treatment group and the control group in the dataset). For illustration, in our study, race is the variable that is suspected to induce bias. Therefore, observations where race is Black or African American are included in the treatment group, and all the other observations are included in the control group. In the meantime, one outcome variable (i.e., mortgage decision) is necessary for the dataset to establish causal inferences of the relationship between the sensitive variable and the outcome variable. All other variables are used as predictors that contribute to the outcome variable. Both predictors and the outcome could be either numerical or categorical.

To establish a causal inference with our data structure, we leverage three bias detection algorithms: (1) Fair on Average Causal Effect (FACE), (2) Fair on Average Causal Effect on the Treated (FACT), and (3) Causal Forest. In particular, FACE and FACT utilize the same background approach. FACE evaluates the disparity of the result by comparing the treatment group observations and the control group observations. It assigns a stabilized weight to each observation and calculates the estimated coefficients of the sensitive variables and other predictors by generating a weighted logistic regression model which involves the sensitive variable, predictors, and the outcome variable. The estimated coefficient of the sensitive variable with the chosen level of significance indicates whether the bias to the treatment group exists. If the coefficient is statistically insignificant, it is considered there is no statistical difference between the treatment group and the control group. In other words, the bias induced by the sensitive variable does not exist in the dataset. Meanwhile, if the coefficient is statistically significant, it represents the causal differences between the treatment group and the control group. For example, in a dataset used in this study, the sensitive variable is race and Black and African American applicants belong to the treatment group, and the outcome is whether the mortgage application is approved or declined. If the coefficient of the treatment group is statistically significant, it indicates that the approval rate is different between Black and African American applicants and other applicants and such a difference is causal. As for FACT, the method first finds the matched pair from the treatment group and the control group by evaluating the values from predictors. It matches a pair of observations with specific matching methods, including Nearest Neighbor Matching (NNM), Nearest Neighbor Matching with a Propensity Caliper (NNMPC), and Mahalanobis Metric Matching within the Propensity Caliper (MMMPC). For illustration, an African American mortgage applicant is matched with a non-African American mortgage applicant who has similar profiles (i.e., their predictors such as income level, credit scores, and property values are similar). Note that observations in a matched pair may not contain the exact same values. As long as the difference is not statistically significant, the matching process is generally considered successful. With matched observations, the causal inference is established by developing a weighted logistic regression model and calculating the coefficient of the sensitive variable. The interpretation of the coefficient is similar to that of FACE. Lastly, as an alternative method of FACE and FACT, Causal Forest implements the random forest algorithm to establish causal inferences. The dataset is split into a tree structure by sampling in order to minimize the prediction errors and to maximize the differences between trees. Causal Forest estimates the average treatment effects to verify whether the sensitive variable is biased to the outcome and the coefficient value represents the impact of the bias on the outcome.

To validate the results from FACE, FACT and Causal Forest, we conduct falsification tests to check the plausibility that the bias detected occurs solely by chance. Specifically, the falsification tests select ten variables and treat them as sensitive variables in the bias detection process through FACE, FACT, and Causal Forest. If we also find a statistically significant coefficient of these variables, it indicates that our main result may not be reliable since bias can be detected from any variables in the dataset. On the other hand, if the coefficient of these variables is not statistically significant, we could conclude that our method is reasonably reliable to establish causal inferences regarding the racial bias in the AI model on mortgage applications.

3.2.2 Detecting Algorithmic Bias with Prediction Accuracy

In addition to detecting bias using causal-based methods, we also employ an alternative bias detection approach that establishes biases based on the difference in prediction accuracy across observations with different sensitive variables. Here, we use two open-source tools: IBM AI Fairness 360 (AIF360) and Microsoft Fairlearn (MSF).

IBM AIF360 included three types of fairness processing: fair pre-processing, fair in-processing, and fair post-processing. AIF360 provides different metrics to detect bias along with the magnitude. For the fairness pre-processing stage, the method inspects the input data that are used to train, validate, and test the AI model. For fairness in-processing, the method examines the trained model to detect bias in the prediction process. For the fairness post-processing, the method evaluates the output predictions and estimates the impact of bias. At the same time, AIF360 also offers a suite of algorithms that can mitigate biases [21]. In our study, we leverage AIF360 to supplement the bias detected by FACE, FACT, and Causal Forest. Particularly, we use AIF360 to detect biases based on the discrepancy of prediction accuracy between the treatment group and the control group. Since the outcome variable in our study is binary (i.e., whether the mortgage application is approved or declined), we calculate an average odds error to represent the absolute False Positive Rate (FPR) and True Positive Rate (TPR) for the control and treatment groups.

Similar to AIF360, Microsoft MSF is another open-source library that facilitates fairness in AI predictions. MSF also provides numerous metrics to assess fairness and algorithms to mitigate the detected bias. The package supports most common fairness metrics including Demographic Parity (DP), Equalized Odds (EO), True Positive Rate Parity (TPRP), False Positive Rate Parity (FPRP), and Error Rate Parity (ERP). These metrics can be used to directly measure biases in the dataset or the model, while they can also be used to evaluate the improvement after the mitigation functions, including exponentiated gradient reduction method, grid search reduction method, threshold optimizer, and correlation remover [19]. We use MSF to detect the bias based on TPRP and FPRP in this study.

3.2.3 Mitigating Bias in the Training Dataset

There are many bias mitigation methods available to the public. In this study, we use the Exponentiated Gradient (EG) bias reduction method available as a part of the Microsoft Fairlearn package [19] to mitigate the bias in AI predictions of mortgage applications. This method utilizes the reduction approach [27] with the EG algorithm [28] to build a fair predictor by reweighting and relabeling based on the Demographic Parity (DP) fairness constraint.

4 RESULTS

4.1 Model-Free Statistics

Before detecting bias by FACE, FACT and Causal Forest, we first provide model-free statistics of our dataset. Specifically, we observe the race distribution in the HMDA 2019 dataset to understand the nature of the data. Some of the races such as American Indian and Asians are minority groups in the United States, so it is expected that we observe lower counts of applications for these races. Table 2 reports the counts of applications with approval and denial decisions and the applicants' races distribution in the HMDA 2019 dataset. In over 2 million observations, 63.74% of the mortgage applications were from White applicants. Meanwhile, the treatment group in our study (i.e., Black or African American applicants) covers 7.67% of the total applications. Note that 20.72% of the applicants did not share their racial information. Except the information unavailable group, White applicants have the highest probability of being approved on their mortgage applications at 57.01%. Meanwhile, the percentage of approval for Black or African American applicants is at 22.14%, which is the second-lowest group after American Indian or Alaska Native.

Table 2: Races distribution and mortgage application final decisions statistics

Races	Count of applications	Percentage of total applications	Count of approval applications	Count of denial applications	Percentage of approval	Average loan amount
Black or African American	219,601	7.67%	39,805	179,796	22.14%	$138,735
American Indian or Alaska Native	40,107	1.40%	6,309	33,798	18.67%	$144,296
Asian	184,922	6.46%	61,550	123,372	49.89%	$304,814
White	1,824,109	63.74%	662,349	1,161,760	57.01%	$204,900
Information not provided by applicant	592,888	20.72%	275,782	317,106	86.97%	$249,847
All	2,861,627	100.00%	1,087,119	1,774,508	61.26%	$214,742

Moreover, for the average loan amounts of different races, Table 2 demonstrates that the average loan amount for Black or African American applicants is the lowest compared to other groups even though it was already difficult for them to obtain their mortgage. Interestingly, the highest average loan amount came from Asian applicants whose percentage of approval was more than twice of that of the probability of approval for Black applicants. These model-free numbers indicate the plausibility of racial bias in the mortgage approval process. Next, we develop an AI model that uses this data as a training dataset and formally detect the bias in the prediction process.

4.2 Bias in AI Predictions

In this subsection, we develop AI models to predict whether each mortgage application should be approved or declined based on the dataset described in the previous subsection. To reduce the computational resources required in the model development process, we randomly select 10,000 observations from the dataset as a subset, develop the model using 10-fold cross-validation for each subset, and repeat this process 100 times and report the average results.

4.2.1 Bias with Causal Inferences

First, we utilize FACE, FACT, and Causal Forest to establish causal inferences in the bias detection process. The estimate of these methods indicates the percentage that the treatment group (i.e., Black or African American applicants) obtains approval from their mortgage applications compared with the control group (i.e., applicants of other races). We also report the p-value, which represents the statistical significance of the estimate. Results are reported in Table 3.

Table 3: Results from FACE, FACT, and Causal Forest

Methods	Average estimate	Average p-value
FACE	-0.45218	0.001247
FACT – NNM	-0.47116	0.000161
FACT – NNMPC	-0.47318	0.019476
FACT – MMMPC	-0.47116	0.000154
Causal Forest	-0.04138	0.000678

The results indicate that the bias exists in the prediction process where the logistic regression model is used. The estimate of all models is negative, indicating that the treatment group (i.e., Black or African American applicants) are less likely to obtain approval for their mortgage applications than the control group. The p-value of all models (except FACT – NNMPC) is lower than 0.001 while that of FACT – NNMPC is lower than 0.05, indicating that the estimate is extremely unlikely to be non-zero by chance.

4.2.2 Bias in Prediction Accuracy

Next, we detect the bias in AI prediction based on an alternative method where the bias is defined as a discrepancy in the prediction accuracy of observations from the treatment group compared with that of observations from the control group.

Using the same dataset, we utilize AIF360 and MSF to develop two models to predict whether an application should be approved or declined. The first model is based on observations in the treatment group (i.e., Black or African American applicants) while the second model is based on observations in the control group (i.e., applicants from other races). The accuracy of the two models is reported in Table 4.

Table 4: Results from AIF360 and MSF

Methods	Average treatment group accuracy	Average control group accuracy
AIF360	92.26%	86.80%
MSF	92.26%	87.13%

As observed, the treatment group model has a higher prediction accuracy than the control group model. To formally test the difference, we conduct the paired t-test to test the difference of the accuracy between the two models. The p-value is lower than 0.01 for both AIF360 and MSF. As a result, we can conclude that there is also a racial bias in terms of prediction accuracy in the AI prediction model for mortgage applications as well.

4.2.3 Bias from AI Model versus Bias in Training Dataset

Our results from the previous two subsections demonstrate the bias in AI predictions. These results are not surprising, since the AI models are trained with a dataset that is well known to be biased against the treatment group (i.e., Black or African American applicants). In this subsection, we extend our results by investigating the magnitude of bias from the AI prediction compared with the bias that already exists in the training dataset.

Here, we rely on three metrics to measure the bias in AI predictions and in the training dataset. First, we calculate the unprivileged ratio, which is the ratio of the positive outcomes (i.e., approved applications) among Black or African American applicants. Second, we calculate the privileged ratio, which is the ratio of the positive outcomes among applicants with other races. Third, we calculate the disparate impact ratio, which is the ratio of positive outcomes in the unprivileged group divided by the ratio of positive outcomes in the privileged group. Intuitively, if Black or African American applicants attain positive outcomes at the same rate as applicants of other races, the disparate impact ratio would be 1, indicating perfect fairness. Meanwhile, the disparate impact ratio of 0 would indicate the other side of the spectrum where the unfairness is at the highest. These three metrics are first calculated using the dataset itself. Then, we build an AI model using the logistic regression algorithm, obtain the prediction results, and calculate these three metrics based on the prediction results. Table 5 reports the findings.

Table 5: Bias in the Training Dataset vs. Bias from AI Predictions

	Bias in the training dataset	Bias from AI predictions
Unprivileged ratio	0.18126	0.11052
Privileged ratio	0.39640	0.33679
Disparate impact ratio	0.45726	0.32817

Consistent with the prior literature, we find that the bias indeed exists in the training dataset since the disparate impact ratio in the training dataset is 0.45726, indicating that Black or African American applicants are approximately 54% less likely to attain positive outcomes than applicants of other races. However, surprisingly, we find that the AI model amplifies this bias by more than 13 percentage points since the disparate impact ratio calculated from the prediction results is 0.32817, indicating that Black or African American applicants are approximately 67% less likely to attain positive outcomes than applicants of other races if the decision is made based on an AI model.

4.2.4 Falsification Tests

Our results so far demonstrate the bias in AI predictions where the race of applicants is a sensitive variable. It is plausible that such a phenomenon occurs by chance, and bias could be detected for any sensitive variable. In this subsection, we conduct falsification tests to alleviate such a concern. Specifically, we select 6 binary variables from the dataset that are unlikely to lead to bias in AI predictions, treat them as a sensitive variable, and rerun our analyses using FACE, FACT, and Causal Forest. The results of falsification tests are reported in Table 6.

Table 6: Falsification Tests

Potential Biases (Column Name)	Column Description	Average Estimate	Average P-value
tract_one_to_four_family_homes	Dwellings that are built to houses with fewer than 5 families.	0.01308	0.3763
ffiec_msa_md_median_family_income	FFIEC Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC).	0.08954	0.2745
tract_population	Total population in tract	0.04439	0.3538
derived_msa-md	The 5-digit derived MSA (metropolitan statistical area) or MD (metropolitan division) code. An MSA/MD is an area that has at least one urbanized area of 50,000 or more population.	-0.00109	0.4692
occupancy_type	Occupancy type for the dwelling	-0.02392	0.4267
other_nonamortizing_features	Whether the contractual terms include, or would have included, any term, that would allow for payments other than fully amortizing payments during the loan term.	0.44349	0.0806

Observe that none of the coefficients of the 6 binary variables that we select for falsification tests is statistically significant at p-value < 0.05, indicating that we cannot statistically conclude that these estimates are non-zero. In other words, we find that these variables, when treated as a sensitive variable, do not generate biased prediction results. Hence, it is unlikely that the bias we detect in our study where Black or African American applicants are less likely to attain approval from their mortgage applications happens by chance.

4.3 Implications of Bias Mitigation in AI Predictions

It is clear that the use of common, off-the-shelf AI models to approve or decline mortgage applicants induces racial bias in the prediction process. This is because the training dataset is historically biased against the treatment group (i.e., Black or African American applicants). Even then, we also find that the use of an AI model amplifies such an existing bias by more than 13 percentage points. In this subsection, we investigate the implications of the use of fair AI methods that are designed to alleviate bias in AI predictions.

Table 7 summarizes the performance of the AI model before bias mitigation, and Table 8 presents the performance of the model with Exponentiated Gradient (EG) debiasing. We report the accuracy and selection rate difference before and after implementing the EG reduction technique to compare if the bias mitigation technique produces fair prediction results.

The original model has the average accuracy of 87.21% and the selection rate of 26.79%. However, note that the selection rate is only 10.73% among the treatment group (i.e., Black or African American applicants) while the selection rate is 28.13% for applicants of other races. The selection rate difference between groups is thus 17.40 percentage points. In the meantime, results in Table 8 demonstrate that, consistent with prior literature, the bias mitigation reduces the overall accuracy of the model by about 3 percentage points. However, the EG bias mitigation process successfully reduces bias in prediction results as evidenced by the reduction of selective rate difference between two groups from 0.1740 to 0.0146. In other words, with the EG bias mitigation algorithm, applicants from the treatment group and those from the control group receive an approval at almost the same rate.

Table 7: Model Performance Before Bias Mitigation

	Accuracy	Selection rate	Count
Overall	0.8721	0.2679	572,326
Treatment group	0.9259	0.1073	44,018
Control group	0.8676	0.2813	528,308
Difference (control group – treatment group)	-0.0583	0.1740	484,290

Table 8: Model Performance After Bias Mitigation

	Accuracy	Selection rate	Count
Overall	0.8439	0.2973	572,326
Treatment group	0.7788	0.2838	44,018
Control group	0.8493	0.2985	528,308
Difference (control group – treatment group)	0.0706	0.0146	484,290

Now that we demonstrate that the bias mitigation technique works to produce fairness in prediction results as intended, we next examine the implications of this imposed fairness. Particularly, we calculate model accuracy, predictive performance, approval counts, approval rates, average approved loan amounts before vs. after bias mitigation.

Table 9: Model Comparison before and after Debiasing

	Accuracy		Predicted Approval Counts		Predicted Approval Rates	
	Black/African American	Other races	Black/African American	Other races	Black/African American	Other races
Fairness-unaware model	92.59%	86.76%	4722	148611	10.73%	28.13%
Biased mitigated model	77.88%	84.93%	12494	157679	28.38%	29.84%
Difference	-14.71% ▼	-1.83% ▼	7772 ▲	9068 ▲	17.65% ▲	1.71% ▲
Truth value			7895	209529	17.94%	39.66%

First, we report model accuracy, predicted approval counts, and predicted approval rates before and after bias mitigation in Table 9. As discussed earlier, the accuracy of the AI model decreases when the fairness constraint is imposed to provide bias mitigation. However, it is worth noting that even though the overall

accuracy decreases by about 3 percentage points, the prediction accuracy for Black or African American applicants decreases by almost 15 percentage points while the prediction accuracy for applicants of other races decreases by almost 2 percentage points only. Meanwhile, because of bias mitigation, the number of Black or African American applicants who are approved for the mortgage increases by about 165% from 4,722 applications to 12,494 applications. In the meantime, applicants of other races also benefit from bias mitigation as the number of applicants who are approved for the mortgage increases by about 6% from 148,611 applicants to 157,679 applicants.

With the increase in the number of approved applications, the natural question that follows is the impact of such a phenomenon on the financial institution. Table 10 and Table 11 report the confusion matrix of the fairness-unaware model and that of the bias-mitigated model respectively.

Table 10: Confusion Table of Fairness-Unaware Model

		Predicted	
		Application denied	Application approved
Truth	Application denied	350,351	4,551
	Application approved	68,642	148,782

Table 11: Confusion Table of the Bias-Mitigated model

		Predicted	
		Application denied	Application approved
Truth	Application denied	333,863	21,039
	Application approved	68,290	149,134

The confusion matrixes of both models confirm that the overall accuracy of the bias-mitigated model decreases. However, it is worth noting that the number of false positive (i.e., applications that are predicted as approved when they should be declined) increases by 16,488 applications (representing 362.29% increases). It is commonly acknowledged that false positive predictions for this type of prediction tasks are costly to the financial institution [29]. As a result, such an increase represents a significant cost of fairness imposed on the financial institution that uses fair AI models to provide fairness in the mortgage application process.

Table 12: Model Comparison on Average Approved Loan Amount

	Average Amount of Loan Approved		
	Truth	Fairness-unaware model	Bias-mitigated model
Treatment group	$221,750	$258,250	$163,820
Control group	$227,750	$310,750	$288,190
Difference (control group – treatment group)	$6,000	$52,500	$124,370

Lastly, we examine the amount of loan that each approved applicant receives. Table 12 reports the average approved loan amount for the fairness-unaware model and that of the bias-mitigated model. The result is particularly interesting. Observe that in terms of the truth value, the average amount of loan approved is very similar ($221,750 for the treatment group and $227,750 for the control group). The

difference is statistically insignificant (i.e., p-value > 0.10) based on the independent t-test. However, when we train an AI model based on such a dataset, the discrepancy in the average amount of loan approved appears. Based on the predictions, the control group receives significantly higher amount of loans (both compared to the true value and compared to the treatment group). Such a discrepancy becomes even worse under the bias-mitigated model. In other words, when the bias-mitigated model is used, Black or African American applicants are more likely to receive an approval decision. However, they receive much less amount of loan.

5 DISCUSSIONS AND CONCLUSIONS

Issues surrounding bias and discrimination in the housing market have attracted interest from both researchers and policymakers for decades. Although several issues in this area, such as bias and discrimination among landlords, have become less prominent in the last several years, some issues, such as bias and discrimination in mortgage approvals, continue to persist. This specific issue is especially important given that the use of AI to assist the approval process has gained traction. In this paper, we empirically examine the implications of the use of AI in the mortgage approval process with respect to bias and fairness. Particularly, we ask the following research questions. (1) Is the use of common off-the-shelf AI models with historically biased datasets amplify the bias? (2) Can the use of fair AI models alleviate the bias that exists in the training dataset? (3) What are the implications of the use of fair AI models?

To answer our research questions, we utilize a dataset that was released through the Housing Mortgage Disclosure Act (HMDA) in 2019. The dataset contains anonymized information on each mortgage application and the approval/denial decision. We use several techniques, including Fair on Average Causal Effect (FACE), Fair on Average Causal Effect on the Treated (FACT), Causal Forest, IBM AI Fairness 360 Suite, and Microsoft Fairlearn toolkit, to empirically examine the dataset. Overall, we find that racial bias indeed exists in the HDMA 2019 dataset. Specifically, Black or African American applicants are less likely to be approved for a mortgage. This result is remarkably consistent across multiple methods that we employ. In addition, we find that racial bias becomes more severe in the prediction results compared to the bias that exists in the dataset. In other words, when an AI model is trained by past decisions to approve/deny mortgage applications, the model captures both the relationship between the factors that are historically used to approve the applications and the racial bias embedded in the process. More importantly, when the trained model is deployed without human intervention, the bias captured by the model is amplified by the model's objective to maximize prediction accuracy. Following that, we implement the Exponentiated Gradient (EG) algorithm, which is a part of the Microsoft Fairlearn toolkit, to mitigate the bias in prediction results. Overall, we find that the bias mitigation works as intended. The selection rate difference (i.e., the difference in the approval rate of Black or African American applicants and that of applicants of other races) is reduced from 17.40% to 1.46%. With this bias mitigation tool, the overall accuracy of the prediction decreases. Meanwhile, the number of mortgage applicants increases by 165% for Black or African applicants and 6% for applicants of other races. Because of this increase, the overall false positive rate of the bias-mitigated model increases by 362%, which may impose a significant burden on financial institutions that adopt the fair AI model. Lastly, even though more Black or African American applicants receive approval decisions for their mortgage applications, the average loan amount significantly decreases by 36%. Our findings demonstrate that the fair AI model may produce fairness in prediction results as defined by the notion of

non-discrimination used in the fairness constraint. However, the implications of the use of fair AI models are far from straightforward. Additional considerations from multiple perspectives, including consumer surplus, social welfare, and legal issues, are still lacking in the literature, as well as empirical results from the real-world implementations of fair AI models. We hope this work opens a conversation on this important issue, especially in the housing market that is known to be prone to bias and discrimination against people of color.

Our work has several limitations, which also represent a great avenue for future research. First, our results on the implications of fair AI are based on a single AI model. Even though other models will most likely produce qualitatively similar results, it would be interesting for a meta-study to compare and contrast results from other fair AI models, especially those that utilize a different fairness constraint. Second, our findings are based on analyses of an archival dataset. Future research that can implement a field experiment would be an excellent addition to the literature. Third, additional research that can follow up on mortgage applicants to observe the eventual outcome of each approved application (i.e., whether it is default or not) would add additional insights into the performance of predictive AI models in terms of business profit and social welfare. Lastly, this study uses the dataset from the United States mortgage market. It would be interesting to examine if there exists a heterogeneity of the implications of fair AI with respect to the geographical region or culture.

CONFLICT OF INTEREST DECLARATION

On behalf of all authors, the corresponding author states that there is no conflict of interest.

REFERENCES

[1] Ross, S. L. and Turner, M. A. Housing discrimination in metropolitan America: Explaining changes between 1989 and 2000. *Social Problems*, 52, 2 (2005), 152-180.

[2] Yinger, J. *Closed doors, opportunities lost: The continuing costs of housing discrimination*. Russell Sage Foundation, 1995.

[3] Wachter, S. M. and Megbolugbe, I. F. Impacts of housing and mortgage market discrimination racial and ethnic disparities in homeownership. *Housing Policy Debate*, 3, 2 (1992), 332-370.

[4] Quillian, L., Lee, J. J. and Honoré, B. Racial discrimination in the US housing and mortgage lending markets: a quantitative review of trends, 1976–2016. *Race and Social Problems*, 12, 1 (2020), 13-28.

[5] Cohen, M. C., Dahan, S., Khern-am-nuai, W., Shimao, H. and Touboul, J. The Use of AI in Legal Systems: Determining Independent Contractor vs. Employee Status. *Queen's University Legal Research Paper Forthcoming*, Available at SSRN: https://ssrn.com/abstract=4013823 or http://dx.doi.org/10.2139/ssrn.4013823 (2022).

[6] Lacruz, F. and Saniie, J. *Applications of Machine Learning in Fintech Credit Card Fraud Detection*. IEEE, City, 2021.

[7] van Esch, P., Black, J. S. and Arli, D. Job candidates' reactions to AI-enabled job application processes. *AI and Ethics*, 1, 2 (2021), 119-130.

[8] Shimao, H., Khern-am-nuai, W., Kannan, K. and Cohen, M. C. Strategic Best Response Fairness in Fair Machine Learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), 664-664.

[9] Favaretto, M., De Clercq, E. and Elger, B. S. Big Data and discrimination: perils, promises and solutions. A systematic review. *Journal of Big Data*, 6, 1 (2019), 1-27.

[10] Schneider, V. Locked out by Big Data: How Big Data Algorithms and Machine Learning May Undermine Housing Justice. *Colum. Hum. Rts. L. Rev.*, 52 (2020), 251.

[11] Ben Shahar, T. H. Educational justice and big data. *Theory and Research in Education*, 15, 3 (2017), 306-320.
[12] Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C. R. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10 (2018), 113-174.
[13] Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. and Podsakoff, N. P. Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of applied psychology*, 88, 5 (2003), 879.
[14] Nisar, Q. A., Nasir, N., Jamshed, S., Naz, S., Ali, M. and Ali, S. Big data management and environmental performance: role of big data decision-making capabilities and decision-making quality. *Journal of Enterprise Information Management* (2020).
[15] Khademi, A., Lee, S., Foley, D. and Honavar, V. *Fairness in algorithmic decision making: An excursion through the lens of causality*. City, 2019.
[16] Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100, 469 (2005), 322-331.
[17] Khademi, A. and Honavar, V. *Algorithmic bias in recidivism prediction: A causal perspective (student abstract)*. City, 2020.
[18] Athey, S. and Wager, S. Estimating treatment effects with causal forests: An application. *Observational Studies*, 5, 2 (2019), 37-51.
[19] Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H. and Walker, K. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).
[20] Das, P., Ivkin, N., Bansal, T., Rouesnel, L., Gautier, P., Karnin, Z., Dirac, L., Ramakrishnan, L., Perunicic, A. and Shcherbatyi, I. *Amazon SageMaker Autopilot: a white box AutoML solution at scale*. City, 2020.
[21] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S. and Mojsilovic, A. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).
[22] Cocheo, S. Justice Department sues tiny South Dakota bank for loan bias. *American Bankers Association. ABA Banking Journal*, 86, 1 (1994), 6.
[23] Dymski, G. A. Why the subprime crisis is different: a Minskyian approach. *Cambridge Journal of Economics*, 34, 2 (2010), 239-255.
[24] Dahan, S. A Path-Dependent Deadlock: Institutional Causes of the Euro Crisis. *Cornell Int'l LJ*, 49 (2016), 309.
[25] Dymski, G. Bank lending and the subprime crisis. *The Handbook of the Political Economy of Financial Crises* (2013), 411.
[26] Bradford, C. *Risk or race?: Racial disparities and the subprime refinance market*. Center for Community Change Washington, DC, 2002.
[27] Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. and Wallach, H. *A reductions approach to fair classification*. PMLR, City, 2018.
[28] Kivinen, J. and Warmuth, M. K. Exponentiated gradient versus gradient descent for linear predictors. *information and computation*, 132, 1 (1997), 1-63.
[29] Larose, D. T. *Data mining and predictive analytics*. John Wiley & Sons, 2015.