Minds and Machines (2021) 31:165–191
https://doi.org/10.1007/s11023-020-09529-4

GENERAL ARTICLE

Algorithmic Fairness in Mortgage Lending: from Absolute Conditions to Relational Trade-offs

Michelle Seng Ah Lee[1] · **Luciano Floridi**[2]

Received: 23 March 2020 / Accepted: 5 June 2020 / Published online: 9 June 2020

Abstract
To address the rising concern that algorithmic decision-making may reinforce discriminatory biases, researchers have proposed many notions of fairness and corresponding mathematical formalizations. Each of these notions is often presented as a one-size-fits-all, absolute condition; however, in reality, the practical and ethical trade-offs are unavoidable and more complex. We introduce a new approach that considers fairness—not as a binary, absolute mathematical condition—but rather, as a relational notion in comparison to alternative decisionmaking processes. Using US mortgage lending as an example use case, we discuss the ethical foundations of each definition of fairness and demonstrate that our proposed methodology more closely captures the ethical trade-offs of the decision-maker, as well as forcing a more explicit representation of which values and objectives are prioritised.

Keywords Algorithmic fairness · Mortgage discrimination · Fairness trade-offs · Machine learning · Technology ethics

1 Introduction

Algorithms are increasingly being used to make important decisions to improve efficiency, reduce costs, and enhance personalisation of products and services. Despite these opportunities, the hesitation around implementing algorithms can be attributed to the risk that the algorithms may systematically reinforce past discriminatory practices, favouring some groups over others on the basis of race and gender and thereby exacerbating societal inequalities. While human decision-making and rules-based

✉ Michelle Seng Ah Lee
michelle.sengah.lee@cl.cam.ac.uk

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

[2] Oxford Internet Institute, University of Oxford and Alan Turing Institute, Oxford, UK

processes are prone to their own unfair biases and inaccurate predictions, algorithms have been subject to higher scrutiny due to their limited transparency and their scalability. A decision based on an algorithm is less transparent or explainable than a rules-based process (e.g. if X, then Y). Whereas a human decision-maker with the same cognitive biases may vary his or her decisions, an algorithmic decision based on a bias is capable of perpetuating discrimination at-scale.

Credit risk evaluation is a highly regulated domain area in the United States and has not yet widely adopted algorithmic decision-making. New financial technology companies have started to push the boundaries by using non-traditional data such as location, payment, and social media to predict credit risk (Koren 2016). Even in mortgage lending, scholars have begun to consider whether machine learning (ML) algorithms can lead to more accurate predictions of default and whether there is an increase in financial inclusion of those who would not have received a loan under a simpler decision-making process (Fuster et al. 2017). There has been a body of established literature on how access to credit is crucial to promoting economic growth, and exclusion from this financial access can trap individuals in poverty without the possibility of upward mobility (King and Levine 1993). Unequal access to credit based on race and gender cripples the previously disadvantaged groups and widens the gap in welfare.

To ensure that algorithmic predictions are "fair," scholars have introduced numerous definitions to formalise fairness as a mathematical condition, such as equal odds, positive predictive parity, and counterfactual fairness (Hardt et al. 2016; Chouldechova 2017; Kusner et al. 2017). This has led to attempts at differentiating between them, such as the tutorial on "21 Definitions of Fairness and their Politics" (Narayanan 2018) and the article "Fairness Definitions Explained" (Verma and Rubin 2018). Others have launched technical implementations of the definitions to produce reports on whether an algorithm passes or fails each test, such as the Aequitas tool by UChicago (Saleiro et al. 2018). The abundance of fairness definitions has obfuscated which definition is most suitable for each use case. It is difficult to interpret the outcomes of the various fairness tests, given that many of them are mathematically incompatible (Kleinberg et al. 2018). This is especially true in mortgage lending, where there has been a documented history of discrimination against black borrowers. Moreover, the drivers of default on a mortgage are not well understood because various borrower and macroeconomic features, many of which are not measured, can contribute to this outcome. These factors preclude the ability to mathematically disentangle the proxies of default risk from proxies of race to equalise the algorithms' predictions for black and white borrowers.

We use the complex sources of discrimination shown in mortgage lending literature as an indicative example to discuss the limitations of existing approaches to fairness at measuring racial discrimination. The US mortgage data are used to demonstrate a new methodology that views fairness as a trade-off of objectives—not as an absolute mathematical condition—but in relation to an alternative decision-making process.

This article is organised into four sections. Section 1 discusses the complex sources of discrimination in mortgage lending, and Sect. 2 introduces the data and algorithms used in this article. Section 3 demonstrates some of the contextual

limitations of existing definitions of fairness. Section 4 proposes a new methodology of trade-offs between objectives.

Overall, we to show that existing approaches to fairness do not sufficiently capture the challenge of racial discrimination in mortgage lending and propose a new methodology that focuses on relational trade-offs rather than absolute mathematical conditions. The US mortgage lending data are used as an example case study to bring to life some of the contextual complexities. It is important to note that it is not the purpose of this article to attempt to draw any empirical conclusions about the presence of discrimination in mortgage lending decisions. This is not possible with the limitations of the public data set, and past studies reviewed in Chapter 1 have already addressed this research question. The main contribution of this article is to build on the past literature on fairness as an absolute notion to propose a new conversation on fairness as trade-offs specific to each context. The current proliferation of fairness definitions is confounding the debate by presenting fairness as a generalised, one-size-fits-all, and absolute goal. Without appropriate translation of the prediction disparity into tangible outcomes, the pass/fail fairness tests do not provide sufficient clarity on which algorithm is the best suited for a decision. This article fills this gap by introducing a methodology that views fairness as a relative notion and quantifies the contextual trade-offs, such that it provides the decision-maker with clear, concrete ways in which each algorithm meets his or her objectives.

2 Discrimination in Mortgage Lending

2.1 Legal Framework for Discrimination

To contextualise the analysis of US mortgage lending data, we will focus our discussion on the US legal framework. The Home Mortgage Disclosure Act (HMDA) requires collection and disclosure of individual-level data for lenders, including race and gender, which allows for greater transparency, accountability, and academic scrutiny in a domain with a history of contentious discriminatory practices. Economists have consistently shown the presence of racial discrimination in various US mortgage markets (Munnell et al. 1996), with the exception of a few who argue that any inequality in outcome is driven by differences in default risk (Ladd 1998).

The controversy around what constitutes unfair lending arises in part because some economists' definitions are narrower than the legal definition of discrimination (Ladd 1998). There are two concepts in legal literature: disparate treatment and disparate impact. Recently, there has been a shift in focus from disparate treatment to disparate impact in legal decisions. According to the US Supreme Court decision *Texas Department of Housing and Community Affairs v. the Inclusive Communities Project*, disparate impact is illegal under the Fair Housing Act regardless of intent (Scalia et al. 2015).

Disparate treatment can be sub-divided into taste-based discrimination and statistical discrimination. The former is the exercise of prejudicial tastes, sometimes forfeiting profit. The latter derives from using the applicant's group (e.g. race) to estimate the credit-worthiness, given the expected difference in statistical distribution of

credit-worthiness between groups. Disparate impact goes beyond the intent to discriminate and focuses on the outcome: any policy or practice that puts one protected group at a disproportionate disadvantage is discriminatory.

In anti-discrimination philosophy, Cass Sunstein writes, "*without good reason*, social and legal structures should not turn differences that are both highly visible and irrelevant from the moral point of view into systematic social disadvantages" [emphasis added] (Sunstein 1994). He leaves room for reasonable justification for outcome disparity, which is reflected in the US legal precedent: the "business necessity clause" in the US that provides an exception when the process is "an appropriate means of ensuring that disparate-impact liability is properly limited" (Scalia et al. 2015).

In reality, however, it is challenging to establish that the outcome is driven by "legitimate" features and not by protected characteristics. Proxies of risk are often intertwined with proxies of protected features. A paired audit study found that minority borrowers were more often encouraged to consider FHA loans, which are considered to be substantially more expensive to finance (Ross et al. 2008). In this case, FHA loan type can be considered both a proxy for race and a proxy for default risk. This challenge is exacerbated by the increase in use of non-traditional proxies of risk, e.g. social media and location data. Algorithms can discover patterns in data that predict the desired outcome "that are really just preexisting patterns of exclusion and inequality," which cannot be resolved computationally (Barocas and Selbst 2016). Finally, often "fairness" presumes that there is a ground truth in each person's risk. However, existing data often represent imperfect information on each individual's risk.

2.2 Sources of Discriminatory Bias

Lenders experience information asymmetry and are limited by the data collected in estimating default risk of mortgage applicants. This section will show the complex ways in which this risk may be over-estimated or underestimated, demonstrating the need for an approach that is more conscious of the contextual sources of unfair bias.

2.2.1 Over-Estimation of Minority Risk

There are three sources of bias that can over-estimate minority default risk: selection bias, disparate treatment, and self-perpetuation of the selection bias. First, there is documented discrimination preceding any data record: in the selection of the institution's service area, in their advertising and marketing strategy (Ladd 1998). The limited outreach in high-minority neighbourhoods reduces the likelihood of their application, resulting in selection bias, in which borrowers with certain characteristics are over-represented in the data. The second source of bias is active disparate treatment, which includes both taste-based discrimination—exercise of prejudicial tastes, sometimes forfeiting profit—and statistical discrimination—the use of the applicant's group (e.g. race) to estimate credit-worthiness. Disparate treatment against black borrowers has been demonstrated in a number of paired audit studies,

in which individuals with comparable profiles but different racial backgrounds inquire about a mortgage and record the advised loan amount, terms, and likelihood of approval. A study in Chicago found that black applicants on average were quoted lower loan amounts and receive less information and assistance in applying for a mortgage (Ross et al. 2008). Finally, the problem of selection bias is exacerbated by the self-perpetuation of the discrimination in the subsequent application stages. As loans by black applicants are denied at a higher rate, the counterfactual of whether or not they would have defaulted had their application been approved is not measured and thus unknown.

Given the resulting limitation on the data set's external validity, lenders relying on the data set as the ground truth risk over-estimating the risk of black borrowers and lending only to applications similar to past successes (i.e. fully repaid loans). It is in the lender's financial interest to expand the volume of credit-worthy borrowers. Building the algorithm on the flawed data set necessarily would produce both discriminatory and sub-optimal result.

2.2.2 Under-Estimation of Minority Risk

Disparate impact goes beyond the intent to discriminate and focuses on the outcome: any policy or practice that puts one protected group at a disproportionate disadvantage is discriminatory unless it is "consistent with business necessity" (Scalia et al. 2015). While this is open to interpretation, several regression studies demonstrated that black borrowers have lower approval rates for mortgage than white Americans, which cannot be explained by other legitimate loan or borrower characteristics. Even after controlling for a comprehensive list of all default risk, default cost, and loan characteristics, a prominent study by the Federal Reserve Bank of Boston found that black and Hispanic borrowers are 8% more likely to be denied a loan than non-Hispanic white borrowers (Munnell et al. 1996). Given this disparity cannot be otherwise attributed to a business reason, it would be considered illegal.

Some scholars have argued that higher denial rates may simply be reflecting the higher risk: because black borrowers' default rates are higher on average than those of white borrowers with similar features, the lenders are not discriminating based on taste (Berkovec et al. 1996). However, this is narrower than the legal definition, as disparate impact is illegal regardless of the motivation. These studies have also been widely critiqued, as it is difficult to control for all possible drivers of default, and the difference in means does not indicate an inherent difference in credit-worthiness attributable to race (Ladd 1998).

Could the default risk be higher for black borrowers? Some have attempted to attribute the difference to the risk derived from financial behaviour and racial differences in wealth and asset preferences. This has been refuted, as when controlling for income and education level, the racial disparities in saving behaviour disappears between white and black Americans (Mariela 2018). However, it is possible to imagine there could be a difference in risk due to social and historical prejudice in related markets, which may impact credit-worthiness in an unmeasurable and unpredictable manner. For example, if discrimination in the job market makes minority incomes more unstable (e.g. more likely to be laid off in a restructuring), race could be the

only possible proxy to measure this risk, which would be higher for black borrowers. While selection bias and disparate treatment may lead a lender to overestimate the risk of minority borrowers, market inequalities may underestimate the risk.

This structural and systematic discrimination is likely to manifest itself in the seemingly "legitimate" features that are used to measure the borrower's likelihood of repayment. It would be myopic to view the potentially higher black default risk in isolation from the intersectional discrimination faced by those already marginalised in society (Crenshaw 1989) and the interconnectivity of gender and racial discrimination (Collins 2002). Indeed, mortgage discrimination based on gender and familial status may vary by the applicant's race (Robinson 2002).

If lenders believe that the true risk is, in fact, lower for black borrowers than can be predicted by the data set, they should actively market to and lend to more black borrowers, expanding its customer base to minorities that are credit-worthy but under-represented in the data set. This may reduce the reported precision of the algorithm in the short-term, but with more information collected on these loans, the data would move to more closely reflect their true risk. As for the risk underestimation, better proxies should be measured to estimate the unknowns, e.g. income volatility, which may be correlated to race on aggregate but may vary within racial groups. This points to two incentives for the lender: (1) increase in market share and (2) more precise prediction of default risk.

2.3 Impact of Algorithms

While racial discrimination in lending has been studied for at least the past century, there has been a more recent increase in implementation of machine learning algorithms to predict credit risk and a corresponding increase in scrutiny on whether or not the predictions are fair. For the purpose of this article, algorithms are defined as a process or procedures that extracts patterns from data. Algorithms discussed in this article for mortgage lending are supervised machine learning models that use past examples to predict an outcome. This excludes rules-based processes, such as a scorecard, which follow a set of explicit rules pre-set by a human decision-maker (e.g. if income > X and if loan amount < y, then approve). Algorithms can be both fairer and more accurate than humans or rules-based processes (Kleinberg et al. 2018); however, the widespread discomfort with the use of algorithms to make decisions derives from the tension between the opportunity to more accurately predict default and the risk of systematically reinforcing existing biases in the data, worsening inequalities at an unprecedented scale. In particular, machine learning algorithms may, without explicitly knowing an applicant's race, be capable of triangulating racial information from a combination of the applicant and loan's features. Algorithms can discover patterns in data that predict the desired outcome "that are really just preexisting patterns of exclusion and inequality," which cannot be resolved computationally (Barocas and Selbst 2016).

While ML and non-traditional data are not currently used in the heavily regulated US mortgage lending market, recent literature has focused on whether ML could benefit both the lenders and the potential borrowers by providing more precise

default predictions. Fuster et al., for example, find that using more complex algorithms (random forest) results in an increase in loans for both white applicants and racial minorities (Fuster et al. 2017).

The introduction of algorithmic decision-making has revived the debate on what it means to unfairly discriminate based on race. Human decisionmaking can be affected by subconscious biases that are challenging to track, and policies and processes can have unexpected impact on minority groups. By contrast, algorithms are inherently auditable in their predictions, and the testing of algorithms to see whether they achieve the desired outcome provides an opportunity for researchers and policy-makers to define and formalise what it means to make a fair decision.

3 Methodology

3.1 Data

It is not within the scope of this paper to perform any empirical analyses on the presence of discrimination in mortgage lending decisions. There is insufficient information in the public data set to draw such conclusions. A sample is drawn from the US mortgage lending data set to bring the analysis to life.

The data used are collected under the Home Mortgage Disclosure Act (HMDA) from all lenders above a size threshold that are active in metropolitan areas (estimated to cover 90% of first-lien mortgage originations in the United States). Note that the data set does not include the final outcome of the loan, i.e. whether or not the individual defaulted on the loan, which is only available within the Federal Reserve, and certain pieces of information, e.g. credit scores, are also undisclosed. While this would limit the internal validity of any analysis assessing the drivers of loan approval, as described above, this should not preclude the demonstration of methodologies. The following steps were taken to facilitate the analysis:

- HMDA data from 2011 are used. Since the borrowers in 2011 experienced an overall increase in house prices, any default risk is likely specific to the borrower rather than reflective of a macroeconomic shock (Fuster et al. 2017);
- The features ‘race’ and ‘ethnicity’ are combined to distinguish between non-Hispanic White applicants and Hispanic White applicants;
- Only black and non-Hispanic White borrowers’ loan data are selected, and other race and ethnicities are removed. Past mortgage studies (Ross et al. 2008; Munnell et al. 1996) have shown the greatest disparity in approval rates and treatment between these two groups;

- Loan records with missing data are removed, given they are from exceptional situations only. Data points are missing for income[1] and location[2] in 3.2% of the data set; and
- 50,000 accepted loans and 50,000 denied loans are randomly sampled without replacement to avoid issues arising from the data imbalance and to facilitate the interpretation of accuracy metrics. The true approval rate is 75.6% in the full data set; however, a representative sample is not needed, as it is not the aim to draw any empirical conclusions about the drivers of approval. The full data set has 92.9% white vs. black, while the sampled data has 90.7% white vs. black.

3.2 Algorithms

The objective of the lender is to predict whether the loan would be accepted or denied. Five algorithms with varying mathematical approaches and complexity were selected:

- Logistic regression (LR) with no regularisation
- K-nearest neighbours classification (KNN), with $K = 5$
- Classification and regression tree (CART)
- Gaussian Naïve Bayes (NB)
- Random forest (RF)

The goal of this article is not to provide a complete survey of algorithms to predict default, but rather, to use several algorithms to provide an indicative result to demonstrate the practicality of a methodology. While this is far from a comprehensive list of algorithms, these five were selected to represent the most variations in mathematical models to compare. CART was selected as a logic-based algorithm with strong interpretability, and RF added an ensemble learning method constructing multiple decision trees through bootstrap aggregation (bagging) to reduce overfitting to the training data. Logistic regression was chosen as another interpretable algorithm but with linear decision boundaries. K-nearest neighbours classification is an unsupervised method relying on distance metric. The Naïve Bayes represents a statistical approach with an explicit underlying probability model. These represent the largest variation of mathematical approaches out of some of the most commonly used classification algorithms (Kotsiantis et al. 2007).

All predictors in the feature set (Appendix) were used, with the exception of race, sex, and minority population of the census tract area. In the US, the Equal Credit

[1] Income is not recorded: (1) when an institution does not consider applicant's income in the credit decision, (2) the application is for a multifamily dwelling, (3) the applicant is not a natural person (e.g. a business), or (4) the loan was purchased by an institution (Source: https://www.ffiec.gov/hmda/glossary.htm).

[2] Property address is not reported: (1) if the property address is unknown, (2) if the property is not located in a Metropolitan Statistical Area or Division in which the institution has an office, (3) if loan is to a small business, or (4) when the property is located in a county with a population of 30,000 or less (Source: https://files.consumerfinance.gov/f/201511_cfpb_hmda-reporting-not-applicable.pdf).

```
LR: 0.554186 (0.004676)
KNN: 0.532029 (0.005771)
CART: 0.570200 (0.005473)
NB: 0.610357 (0.005579)
RF: 0.639843 (0.005599)
```

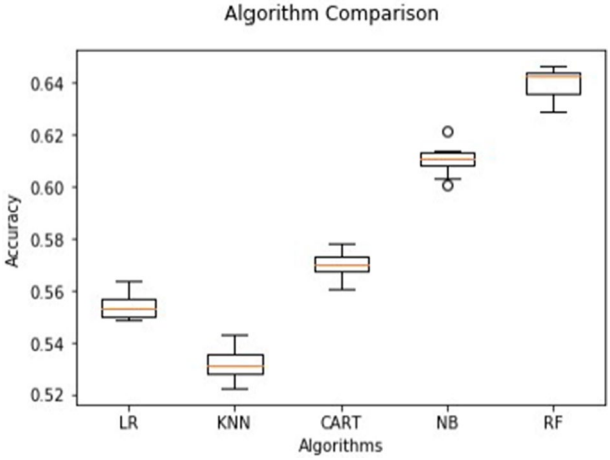

Fig. 1 Algorithm comparison in predicting loan approval

Opportunity Act (ECOA) of 1974 prohibits credit discrimination on the basis of race, and this would preclude lenders from including these features in the decision-making process. Legally, lenders are required to make lending decisions "as if they had no information about the applicant's race, regardless of whether race is or is not a good proxy for risk factors not easily observed by a lender" (Ladd 1998). There are important studies showing that the inclusion of these features may result in both fairer and more accurate outcomes (Kleinberg et al. 2018), which will be addressed in the Sect. 4.4. With tenfold cross validation, the accuracy and standard deviation are shown in Fig. 1.

4 Limitations of Existing Fairness Literature

To address the increasing concern that algorithms may reinforce unfair biases in the data, scholars have introduced dozens of different notion of fairness in the past few years. This sudden inundation of definitions prompted some researchers to attempt to disentangle the nuanced differences between them, such as the tutorial on "21 Definitions of Fairness and their Politics" (Narayanan 2018) and the article "Fairness Definitions Explained" (Verma and Rubin 2018). However, they only focus on the mathematical formalisation of each definition without addressing the broader practical and ethical implications. Given the considerable disagreement among people's perception of what "fairness" entails (Grgic-Hlaca et al. 2018), the absolute representation of fairness oversimplifies the measurement of unfair discrimination. This section categorises fairness definitions into four broad groups and discusses

their approach, ethical grounding, and limitations in the context of mortgage lending.

4.1 Ex Post Fairness

Aligned to the legal rulings that focus on disparate impact, an ex post approach appraises fairness based on the final outcome only, rather than intent or expectation at the time of decision-making.

4.1.1 Group Fairness

Group fairness, or demographic parity, is a population-level metric that requires equal proportion of outcome independent of race. This can be in absolute numbers (in HMDA data, if 1000 black and 1000 white applicants' loans are approved) or in proportion to the population (e.g. of 9350 black and 90,650 white applicants, 935 black and 9065 white applicants' loans are approved). Formally, with $\hat{Y}$ as the predicted outcome and A as a binary protected attribute:

$$P\left(\hat{Y}|A=0\right) = P\left(\hat{Y}|A=1\right) \tag{1}$$

With 29% approval rate for black borrowers and 52% approval rate for white borrowers, the data set is in clear violation of the group fairness metric. It is important to highlight that in the US, the definition of disparate impact is a variation of group fairness. The US Equal Employment Opportunity Commission (EEOC) imposed a rule that the selection rate of the protected group (black borrowers) should be at least 80% of the selection rate of the other group (white borrowers) (Feldman et al. 2015). Given the lenders in this sample data set have approved 52% of loans by white applicants and 29% of loans by black applicants, the black-to-white approval ratio is 55.8%, which is below the 80% threshold and therefore classed as having disparate impact. Feldman et al. (2015) have formalised the approach to identifying disparate impact, but their methodology for pre-processing the data to remove the bias has shown instability in performance of the technique (Feldman et al. 2015).

This rather arbitrary threshold mandates equal outcome irrespective of relative credit-worthiness or differences in financial abilities, which is fundamentally at odds with the attempt to make the best prediction of default risk. When there are features that are causally and strongly associated with outcome (e.g. income) but also strongly correlated to protected group membership (e.g. race/gender), this criterion would actively attempt to correct an existing bias, even if it results in reverse discrimination and loss in accuracy in predicting risk.

4.1.2 Equalisation of Evaluation Metrics

Equalisation of error metrics represent the notion of equality in opportunity, in contrast to equal outcome (group fairness). Each of these attempts to formalise the belief that the predictive accuracy should not vary between racial groups.

Table 1 Confusion matrix

	predicted default	**predicted repay**	**total**
default	True Positive	False Negative	defaulted
repay	False Positive	True Negative	repaid
total	predicted defaulters	predicted non-defaulters	

The lender would select a metric based on the error that matters the most. The False Positives (FP) represent lost opportunity (predicted default, but would have repaid), and the False Negatives (FN) represent lost revenue (predicted repayment, but defaulted). Note that in the mortgage lending use case, the True Positive (TP) and False Positive (FP) rates are unknown. If the borrower is predicted to default, the loan would likely be denied. This problem persists in other domains: the potential performance of an individual who was not hired and the recidivism risk of a criminal who was not granted bail are both unknown.[3]

First, the following are the error rates used in the metrics, with the examples from mortgage lending provided in the Confusion Matrix in Table 1:

- True Positive Rate (TPR) = TP/(TP + FN)
- True Negative Rate (TNR) = TN/(FP + TN)
- False Positive Rate (FPR) = FP/(FP + TN) = 1 − TNR
- False Negative Rate (FNR) = FN/(FN + TP) = 1 − TPR
- Positive Predictive Value (PPV) = TP/(TP + FP)

Some of the most commonly cited fairness definitions in this category include:

- *Equal opportunity/false negative error rate balance* (Hardt et al. 2016) Equal FNR. Among applicants who are credit-worthy and would have repaid their loans, both black and white applicants should have similar rate of their loans being approved;
- *False positive error rate balance/predictive equality* (Chouldechova 2017): Equal FPR. Among applicants who would default, both black and white applicants should have similar rate of their loans being denied;

[3] Selection bias limits the lender's ability to calculate these metrics. To counteract this, some scholars (Verma and Rubin 2018) have used a credit scoring data and defined FN, for example, as the probability of someone with a good credit score being incorrectly being assigned a bad credit score. While this does not reflect the decision-making process in lending, as credit scores would have been provided by a third party (Fair Isaac Corporation—FICO credit scores), it is a useful circumvention of the missing information.

- *Equal odds* (Hardt et al. 2016) Equal TPR and FPR, meeting both of the above conditions;
- *Positive predictive parity* (Chouldechova 2017) Equal PPV. Among credit-worthy applicants, the probability of predicting repayment is the same regardless of race;
- *Positive class balance* (Kleinberg et al. 2016): both credit-worthy white and black applicants who repay their loans have an equal average probability score; and
- *Negative class balance* (Kleinberg et al. 2016) both white and black defaulters have an equal average probability score.

Below are the results, with a heatmap of the within-column values.

4.1.3 Fairness Impossibility

The higher disparity represents a more flagrant violation of the mathematical condition. Consider the logistic regression (LR) and random forest (RF) models. Fuster et al. found, using the full HMDA 2011 data, that RF returned greater outcome disparity between black and white loan applicants than LR (Fuster et al. 2017). This is mostly supported in Fig. 2, as RF has higher disparity in error rates for all metrics except PPP. PPP condition is violated if the probability of predicting repayment (and thus loan approval) is unequal between white and black applicants who would have repaid the loan.

This is as expected, given that equalised odds and PPP are mathematically incompatible (see (Chouldechova 2017) for the proof). In addition, it has been proven impossible to simultaneously satisfy group fairness, negative class balance, and positive class balance (Kleinberg et al. 2016). While the trade-offs between other combinations of definitions have not been formally tested, it is clear that no model would meet all the conditions.

Existing attempts at operationalising fairness tests, such as AI 360 Fairness by IBM (Bellamy et al. 2018) and Aequitas Project by UChicago (Saleiro et al 2018), produce reports with pass/fail results of a combination of the above metrics by using an error threshold (e.g. parity of less than 10%). Without consideration of how the disparity translates into the use case, the thresholds set are rather arbitrary.

Model	Equality of Opportunity (EOP)	False Positive Error Rate Balance (FPERB)	Equal odds (EO)*	Positive Predictive Parity (PPP)	Positive Class Balance (PCB)	Negative Class Balance (NCB)
LR	4%	4%	4%	33%	2%	3%
KNN	6%	6%	6%	18%	8%	8%
CART	6%	6%	6%	21%	16%	18%
NB	7%	7%	7%	44%	41%	41%
RF	9%	9%	9%	3%	18%	18%

* EO condition is met if EOP and FPERB are both met.

Fig. 2 Heatmap table: results of fairness tests (% difference)

Given that many of these metrics are mathematically incompatible, it is difficult to interpret the trade-offs between the fairness conditions. Is a 30 percentage point decrease in positive predictive parity (LR to RF) preferable to a 15 percentage point decrease in negative class balance (RF to LR)? It is difficult to derive meaning from these absolute conditions of fairness.

4.1.4 Proxies of Race and Proxies of Risk

Scholars have recently attempted to map some of these metrics to the moral and ethical philosophical frameworks (Heidari et al. 2018; Lee 2019) to aid in the metric selection process. However, Heidari et al. make a critical assumption that there exists a clear delineation between features that are "irrelevant," e.g. race, and those that can be controlled by the individual. The mortgage market does not exist in a vacuum; while potential borrowers can improve their credit-worthiness to a certain extent, e.g. by building employable skills and establishing a responsible payment history, it is difficult to isolate the features from discrimination in related markets, historical inequalities, and the impact of their personal history. It is challenging to imbue each definition with ethical values when the definitions do not meet the core assumptions of the moral reasoning (Lee 2019).

4.1.5 Existing Structural Bias

One key limitation of these fairness metrics is that they fail to address discrimination already in the data (Gajane 2017). In another piece of work on "fairness impossibility," Friedler et al. point out that each fairness metric requires different assumptions about the gap between the observed space (features) vs. the construct space (unobservable variables). They conclude from their analysis that "if there is structural bias in the decision pipeline, no [group fairness] mechanism can guarantee fairness" (Friedler et al. 2016). This is supported in a critique of existing classification parity metrics, in which the authors conclude that "to the extent that error metrics differ across groups, that tells us more about the shapes of the risk distributions than about the quality of decisions" (Corbett-Davies and Goel 2018). In mortgage lending, with documented structural discrimination, these group-level metrics fail to address the bias already embedded in the data. This problem is shared by the ex ante fairness definitions (Gajane 2017).

4.2 Ex Ante Fairness

Some scholars have introduced metrics to disaggregate fairness, which test for each individual prediction. This is practical for live and continuous decisionmaking, as it returns whether a single decision is fair at that point in time.

4.2.1 Individual Fairness

Various measures of individual fairness have been introduced, in which similar individuals should receive similar outcomes (Dwork et al. 2012). Formally, for similar individuals i and j:

$$\hat{Y}\left(X_{(i)}, A_{(i)}\right) \approx \hat{Y}\left(X_{(j)}, A_{(j)}\right) \tag{2}$$

The challenge of this approach is: how to define "similarity" that is independent of race (Kim et al. 2018). Given minority borrowers were more often encouraged to consider FHA loans, which are considered to be substantially more expensive to finance (Ross et al. 2008), while the type of loan is an important consideration in the probability of default, it is also a partial proxy for race. When the predictive features are also influenced by protected features, designation of a measurement of "similarity" cannot be independent of those protected features. Some scholars have attempted to incorporate active corrections for racial inequality into metrics of similarity (Dwork et al. 2012), but this depends heavily on the assumption that the inequality due to racial discrimination can be isolated from other sources of inequality.

4.2.2 Counterfactual Fairness

Counterfactual fairness condition is whether the loan decision is the same in the actual world as it would be in a counterfactual world in which the individual belonged to a different racial group (Kusner et al. 2017). This metric posits that given a causal model (U, V, F) with a set of observable variables V, a set of latent background variables U not caused by V, and a set of functions F, the counterfactual of belonging to a protected class is independent of the outcome. Where X represents the remaining attributes, A represents the binary protected attribute, Y is the actual outcome, and $\hat{Y}$ is the predicted outcome, formally:

$$P\left(\hat{Y}_{A\leftarrow a}(U) = y | X = x,\ A = a\right) = P\left(\hat{Y}_{A\leftarrow a^0}(U) = y \middle| X = x,\ A = a^0\right) \tag{3}$$

While this provides an elegant abstraction of the algorithm, the causal mechanisms of a default on a mortgage is not well understood. It is also difficult to isolate the impact of one's race on the risk of default from the remaining loan and borrower features. In this particular use case, pre-defining a causal model is especially challenging.

4.3 Limitations in Existing Approaches to Fairness

Existing approaches to fairness impose an absolute mathematical condition, whether comparing outcomes between groups, similar individuals, or group error rates. They do not account for bias embedded in the data. While these metrics are useful in quantifying the disparity, they are agnostic to the context. They have limited interpretability of outcomes specific to each use case and are challenging to set up when the proxy for outcome is intertwined with the proxy for race. Moreover, they only focus on the risk for the minority group without an eye to the potential benefits of the algorithm.

Rather than evaluating fairness against an absolute target (e.g. outcome disparity of 0), a decision-maker may consider fairness as a relative notion, shifting the focus from the quantification of unfairness against a formula to how the algorithm performs in comparison to other possible algorithms. If the goal is to select the best possible algorithm, the fairness metrics may not fully embody what is "best." In reality, a decision-maker has several competing objectives.

While it may be impossible to remove all bias, it is possible to build an algorithm that is fairer and more accurate than the alternative. A new approach is required that builds on the previous fairness definitions but leverages the context-dependency of the data available and the relational nature of whether a decision is fair. This would provide more holistic and tangible information to the lender.

5 Proposal of Trade-off Analysis

This section will propose an alternative approach that can address the contextual complexity and provides a more precise proxy for what represents the "best" algorithm for the decision-maker. This provides information that cannot be derived from the fairness analysis by assessing an algorithm's effectiveness in achieving concrete objectives specific to the domain area. This section will demonstrate this methodology using the US mortgage lending data set.

5.1 Operationalisation of Variables

Two of a lender's objectives that also intersect with public interest are discussed: increasing financial inclusion and lowering denial rate of black borrowers. An increase in access to credit represents the lender's growth in market share, as well as being a global development goal due to its importance for the economy and for individuals' upward mobility (Demirguc-Kunt and Klapper 2012; King and Levine 1993). This is affected by the accuracy of the algorithm in predicting default (i.e. potential revenue loss). The impact on racial minorities is important to consider to comply with regulatory requirements, to manage reputational and ethical risks, and to mitigate the racial bias embedded in the data.

5.1.1 Financial Inclusion

In order to estimate the impact of an algorithm on financial inclusion, the following assumptions are made:

Assumption 1 HMDA data represent the perfect model. In other words, the loans accepted by the HMDA lenders were repaid in full, and the loans denied would have defaulted. Note that in reality, we would not know whether those who are denied a loan would have defaulted; however, this information is not used in this analysis (see definition of negative impact on minorities in Sect. 4.1.2).

Assumption 2 There is one hypothetical lender. The HMDA data represent a multitude of US-based lenders, but this analysis will focus on a lender-level analysis. Future work can assess the impact on a market-level in a multi-player model.

Assumption 3 The lender has a capital limit of $1 billion and gives out loans through the following process:

1. The lender uses an algorithm to predict whether or not the loan will default;
2. The lender sorts the loans by the highest probability of repayment; and
3. The lender accepts the loans with most certain repayments until the capital limit has been reached.

Assumption 4 The loans are either fully paid or will default, i.e. the lender gets the full amount back, or none. This simplifies the expected return calculation by ignoring the differential interest rates and partial payments.

Assumption 5 The lender aims to maximise the expected value of the loan, which is the *accuracy of the algorithm* × *the loan amount*. In other words, if the accuracy of the algorithm is 60% in predicting default, and the loan is for $1 million, then the expected value is $600,000 given there is a 60% chance of full repayment vs. default.

Assumption 6 There is no differentiation in terms, conditions, and interest rates between racial groups. Different rates can be used to pricediscriminate, resulting in unequal distribution of the benefits of financial inclusion. While this is an important consideration for future studies, this analysis will only consider the aggregate-level financial inclusion.

With the above assumptions, financial inclusion can be roughly estimated with the expected value of the loans, which, when maximised, represents the lender's ability to give out more loans. A total expected value of $600 million represents 4000 loans of $150,000 (the median loan value in the data set).

This is a rough definition from a single lender's perspective. While the multifaceted macroeconomic definition and measurement of financial inclusion have been contentious, the primary objective for building an inclusive financial system is to minimise the percentage of individuals involuntarily excluded from the market due to imperfect markets, including incomplete/imperfect information (Amidži et al. 2014). The reduction of portfolio risk with more complete information on each loan's credit-worthiness can be viewed as reducing this asymmetry and improving efficiency. Future work may revisit this definition to expand beyond the lender-level to a market-level and focus on the improved access specifically for low-income and high-risk applicants.

5.1.2 Negative Impact on Minorities

Moving away from what is fair, what is the comparative adverse impact on black mortgage applicants for each algorithm? This will be measured simply as the percentage of black applicants who were denied a loan under the algorithm, which is *the number of denied loans by black applicants/total number of black applicants*. Note this does not consider whether those who were denied a loan would have defaulted. Alternative objectives may be constructed from the fairness metrics, e.g. black-white outcome disparity (group fairness). To demonstrate this as a supplement to the fairness metrics, we will use this raw measure of impact on potential black borrowers.

5.2 Trade-off Analysis

With the variables operationalised as per above, Fig. 3 shows a chart of financial inclusion vs. negative impact on minorities. The baseline represents the outcome at 50% accuracy (random chance) in predicting default. The error bars around each expected value corresponds to the standard deviation in the accuracy of the algorithm in a tenfold cross-validation.

Note that the ordering of the expected value of the loans coincides with the algorithms' accuracies. Some algorithms are better at predicting the loan outcome than others, yet the existing fairness metrics overlook the opportunities and customer benefit provided by a more accurate predictor, which needs to be considered in any evaluation of an algorithm. If the true relationship between default and the input features were linear, a linear model would return the best accuracy. However, given that defaulting on a loan is a complex phenomenon to model with an unknown combination of potential causal factors, it is reasonable to expect that its prediction will be better modelled by more complex (non-linear) algorithms.

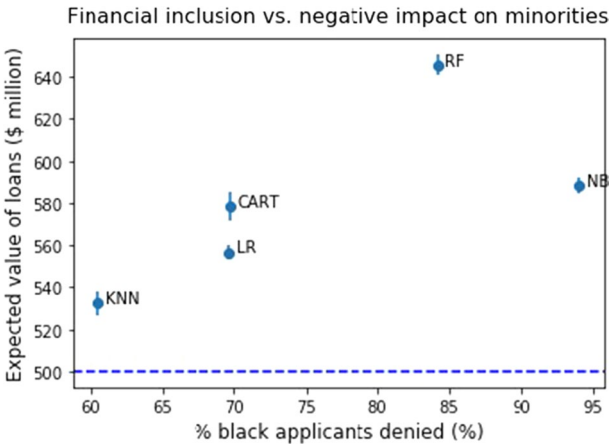

Fig. 3 Trade-off: financial inclusion vs. negative impact on minorities

Also note that overall (e.g. KNN, LR, RF), negative impact on black applicants increases with higher algorithm accuracy. In other words, the increase in the aggregate benefit (financial inclusion) tends to be at the expense of the welfare of the minority and disadvantaged group. There is one notable exception: NB has a much higher denial rates of black applicants than RF, even though its accuracy is lower. This could suggest that NB may be overfitting to the proxies of race. This aligns with the finding of Fuster et al. using the HMDA 2011 data: in a lending system using random forest, the "losers" who would have received a loan under a logistic regression algorithm who no longer qualify are predominantly from racial minority groups, especially black (Fuster et al. 2017).

Random forest is better in absolute terms (in both financial inclusion and impact on minorities) than Naïve Bayes. Random Forest has 10.7 lower percentage points in denial rates for black applicants. Given the median loan value is $150,000, moving from NB to RF results in 368 new medianvalue loans totalling $55.27 million. The decision is more ambiguous between CART and LR. While CART is more accurate and results in greater financial inclusion (equivalent of $15.6 million of loans, or 103 median-value loans), CART results in a 3.8 percentage points increase in denial rates for black loan applicants compared to LR.

This analysis reveals and quantifies the concrete stakes for the decisionmaker, which provides additional information to the absolute fairness tests. It would enable the lender to select the algorithm that best reflects its values and its risk appetite. Of course, not all competing objectives can be quantified; regulatory/legal requirements and explainability of algorithms should all also be considered. For example, RF may be deemed unacceptable due to the relative challenge in interpretability compared to LR. This gap may be narrowing, however, as important progress has been made in recent years in developing model-agnostic techniques to explain machine learning algorithms' predictions in human-readable formats (Ribeiro et al. 2016; Wachter et al. 2017). While these methods are not without their challenges in bridging the gap between real-life and machinelearning objectives (Lipton 2018), they could be used to help interrogate the drivers of each individual prediction regardless of model complexity.

One of the benefits of this methodology is its flexibility. The axes can be adapted to the domain area and the decision-maker's interests. In other use cases, e.g. hiring or insurance pricing, the trade-off curve would look very different. Other algorithms—whether rules-based process or a stochastic model—can also be mapped on this trade-off chart.

5.3 Proxies of Race

Why would some algorithms affect black applicants more than others? Fuster et al. have shown that some machine learning algorithms are able to triangulate race from other variables and features in the model, reinforcing this racial bias (Fuster et al. 2017). This phenomenon is termed "proxy discrimination" (Datta et al. 2017).

When a logistic regression algorithm is run to predict race instead of loan outcome with the same set of features, all features are statistically significant in their associations with race except for: FSA/RHS-guaranteed loan type, owner occupancy status, and property type. The following features are associated with the applicant's race being black: having a lower income and a lower loan amount and being from a census tract area with low median family income, smaller number of 1–4 family units and owner-occupied units, larger population size, and a lower tract-to-MSA/MD income. Of the categorical features, one feature with especially high regression coefficient is the FHA-insured loan type.

With all other categorical features set as the baseline and given the data set's median values for each of the continuous features, the probability of the applicant's race being black if his loan type is FHA-insured is 36.16%. Otherwise, the probability that he/she is black is 19.38%—a 16.78 percentage point difference. This supports the finding from the paired audit study (Ross et al. 2008) that black applicants tend to be referred to the more expensive FHA-insured loans.

Loan type is also statistically significant in a logistic regression to predict loan outcome, with FHA-insured loan type being negatively associated with loan approval probability. Therefore, having an FHA-insured loan type is both a proxy for risk (given the higher cost) and a proxy for race. Given the statistical significance of these features, it is clear that the applicant's race can be inferred to a certain extent from their combination.

This discredits the existing "fairness through unawareness" approach of lenders (Ladd 1998), which attempts to argue that non-inclusion of protected features implies fair treatment. Some have suggested removing the proxies as well as the protected characteristics (Datta et al. 2017). The attempt to "repair" the proxies through pre-processing the data to remove the racial bias has been shown to be impractical and ineffective when the predictors are correlated to the protected characteristic; even strong covariates are often legitimate factors for decisions (Corbett-Davies and Goel 2018). There is no simple mathematical solution to unfairness; the proxy dependencies must be addressed on the systemic level (Ramadan et al. 2018). In the case of mortgages, lenders must review their policy on why an FHA-insured loan type may be suggested over others and how this may affect the racial distribution of loan types.

In addition, recall that one of the key limitations of both ex post and ex ante fairness metrics is their inability to account for discrimination embedded in the data (Gajane 2017). Given that the proxy for risk cannot be separated from the proxies for race, individual fairness metrics would be challenging to set up. With the proof of past discrimination embedded in the data through the marketing, paired audit, and regression studies, mortgage lending fairness also cannot be evaluated through simple disparity in error rates. Having a perfect counterfactual would be desirable, as it would disentangle the complex dependencies between covariates if we had the true underlying causal directed acylic graph. However, this is challenging to achieve in predicting default. For this use case, the trade-off-driven approach is more appropriate than the fairness metrics.

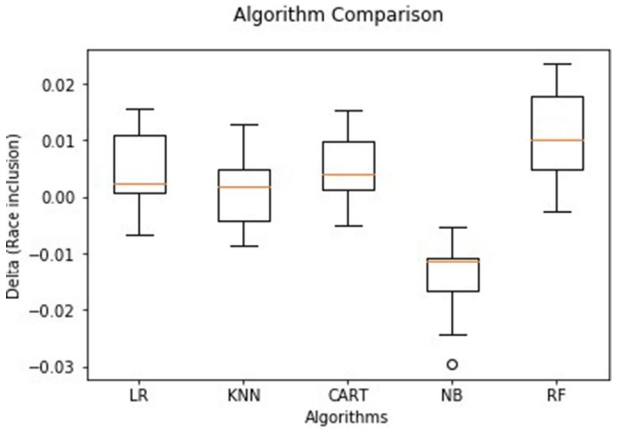

Fig. 4 Change in algorithm accuracy after inclusion of race

5.4 Triangulation of Applicant's Race

If race can, in fact, be triangulated through the remaining loan and borrower characteristics, this implies: (1) the inclusion of race in the algorithm would likely not make a difference in the algorithm's accuracy, and (2) the extent to which the accuracy changes with the inclusion of race depends on the algorithm's ability to triangulate this information. To test this, we included the race in the features to predict loan outcome and re-ran each of the algorithms. With δ_{ij} as accuracy of algorithm j on sample i with race, the changes due to inclusion of race are plotted in Fig. 4.

Two-sample paired t-tests were run on the accuracies before and after the inclusion of race, given they are from the sample cross-validation samples with an added predictor. The results are below with * next to those that are statistically significant at a 5% level:

- LR: t-statistic = 1.71, p-value = 0.12
- KNN: t-statistic = 0.56, p-value = 0.59
- CART: t-statistic = 2.41, p-value = 0.040*
- NB: t-statistic = − 6.22, p-value = 0.00015*
- RF: t-statistic = 4.04, p-value = 0.0029*

There is minimal difference in most algorithms' accuracies. NB and KNN's accuracies are, in fact, worsened by the inclusion of race, but only NB's change is statistically significant. Inclusion of race in CART, LR, and RF positively affect the accuracy in predicting the loan outcome, with the results of CART and RF being statistically significant. NB results may be unstable given that its assumption that the predictor variables are independent is violated, and including race in the feature set results in the algorithms over-fitting to race. RF is better at handling feature dependencies and avoiding over-fitting through its bootstrapping methods (Kotsiantis et al. 2007), and its robustness to redundant information may explain this result.

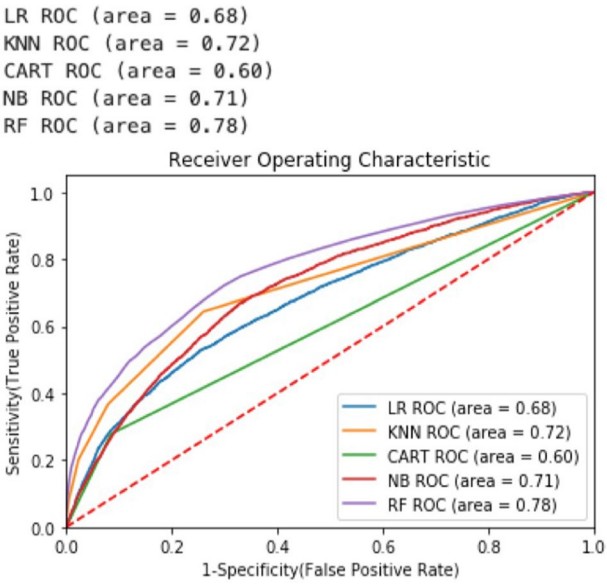

Fig. 5 ROC curves of each algorithm to predict race

If, in fact, racial information is embedded in the other features, can algorithms predict race? Each algorithm was run to predict race based on the given features, rather than loan outcome. Given the imbalance in the race, instead of accuracy metrics, the performance of the algorithms would be best evaluated by the Receiver Operating Characteristic (ROC) and its Area Under the Curve, which measure the True Positive Rate and the False

Positive rate of the algorithms The ROC curves are plotted in Fig. 5. The ideal ROC curve hugs the upper left corner of the chart, with AUC close to 1. The red dotted line can be interpreted as random chance. In this particular data sample, it appears that RF, KNN and NB have relatively higher AUC than LR and CART, showing that they are better performers in predicting race with the given set of features. Further study is required to understand what types of mathematical models are better able to predict protected characteristics and the corresponding impact on the results.

Given these outcomes, it is reasonable that the trade-off curve does not shift much with the inclusion of race. Figure 6 shows the impact of adding race on the trade-off curve. While it does tend to increase the denial rates of black applicants, the difference is small compared to the impact of algorithm selection.

The applicant's race can be triangulated from the remaining features, highlighting the importance of algorithm selection, which has a greater impact on the trade-off between objectives than the inclusion of race as a predictor. The indicative results show that some algorithms are better at triangulating race than others, which should be explored in future studies to help inform the role of model type on the trade-offs.

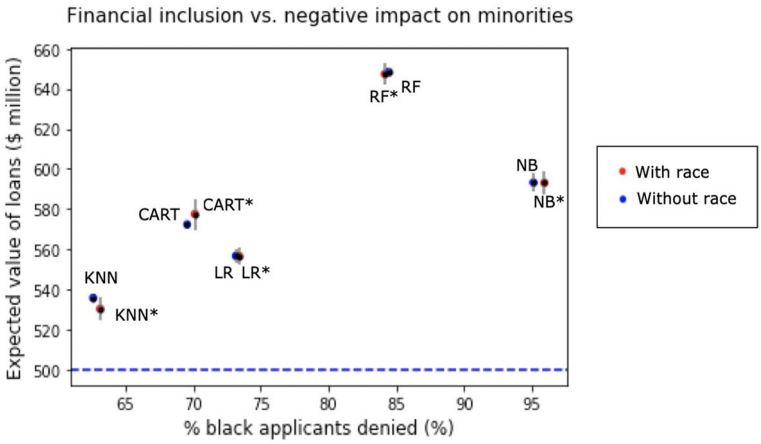

Fig. 6 Trade-off with and without race

In summary, the trade-off can be set up by: (1) defining and quantifying the real-world objectives into measurements, (2) building algorithms and computing relevant metrics, (3) identifying potential proxies of protected characteristics to interrogate whether their inclusion in the model is justifiable, and (4) selecting the algorithm that best reflect the prioritisation of competing objectives.

Overall, the trade-off analysis reveals concrete and measurable impact of selecting an algorithm in relation to the alternatives. Given the ability of algorithms to triangulate race from other features, the current standard approach of excluding race is shown to be ineffective. The disentanglement of proxies for race and proxies for loan outcome, as demonstrated through the analysis of FHA-insured loan types, challenges the assumptions of the existing absolute fairness metrics. When there are multiple competing objectives, this approach can provide actionable information to the lender on which algorithm best meets them.

5.5 Limitations and Future Work

This article aimed to demonstrate the trade-off methodology within the scope of one case study: racial bias in US mortgage lending. The analysis was limited by the data set, which did not provide the full set of information required to mimic the decision-making process of a lender. With additional information on default outcome, terms and conditions of the loans, credit history, profitability of the loan, etc., a more empirically meaningful assessment would be possible of the algorithms and their effectiveness in meeting the lender's goals. Despite the incomplete data, the methodology demonstrated that it is possible to expose an interpretable and domain-specific outcome in selecting a decision-making model. The assumptions made in the operationalisation of variables to compensate for the missing information are mutable and removable.

This paper only begins to unravel the possibilities of the relativistic tradeoff technique, in contrast to the existing approaches to fairness in literature. Some of the areas for future exploration include:

The change in trade-off curve based on different joint distributions between race and default: It would be interesting to visualise the changes in the trade-offs depending on the amount of outcome disparity between black and white borrowers in the data set. The degree to which the increase in aggregate benefit is at the expense of the protected group is likely related to the joint distribution between the predicted outcome and race.

Addition of other algorithm types: It is important to better understand what mathematical set-up leads an algorithm to be more affected by bias in the data. In addition, algorithms that are post-processed to calibrate the predictions, while critiqued in their appropriateness and efficacy (Corbett-Davies and Goel 2018), can be added to the trade-off analysis to examine how they impact the denial rates and financial inclusion.

Generalisation to other domain areas, e.g. pricing, hiring, and criminal recidivism: Depending on the different competing objectives in other domain areas, the variables may change in the trade-off analysis. For example, two of the objectives in a hiring algorithm may be increase in diversity (i.e. percentage of minority groups hired) and an increase in performance metrics of the team. While the technique may be generalisable to other case studies, it would be useful to identify the nuances in the differences in underlying data sets, ethical considerations, and legal and regulatory precedents.

Generalisation to other local markets While this analysis is limited to the US context, further work should contextualise the methodology to local regional contexts, including any documented history of discriminatory practices against legally protected groups and competing priorities in market.

Multi-player model Only one lender's perspective was considered; regulators and policy-makers would be interested in the market-level analysis in which all the lenders' decisions are aggregated. While it was assumed that the lender has the sole authority to decide on which algorithm is the most appropriate, this may be limited by the perspectives of customers and of the market regulators. To understand the policy implications, multiple stakeholders' objectives must be considered.

6 Conclusion

The one-size-fits-all, binary, and axiomatic approaches to fairness are insufficient in addressing the complexities of racial discrimination in mortgage lending. We aimed to introduce and demonstrate a new methodology that addresses the limitations of existing approaches to defining fairness in algorithmic decision-making. The new analysis views fairness in relation to alternatives, attempting to build a model that best reflects the values of a decision-maker in the face of inevitable trade-offs.

There has been a massive proliferation of fairness definitions with only minor variations, without a corresponding unpacking of the nuances in the assumptions and ethical values embedded in the choice of each definition. Given the "fairness impossibility" of many of the formalisations, despite the importance of identifying the most suitable definition for a use case, the focus on defining fairness in a generalisable mathematical formula renders the results of the tests difficult to interpret.

To clarify what is at stake, an approach more attentive to the complexities of the use case could complement the fairness tests by providing more concrete information to the decisionmaker.

We introduced a new methodology that views fairness—not as an absolute mathematical condition—but as a trade-off between competing objectives in relation to alternatives. The goal of the decision-maker is to build an algorithm that is better on multiple dimensions in relation to any existing process or model. By mapping the trade-off between financial inclusion and impact on black borrowers, a lender has access to tangible and explainable justification for selecting one algorithm over another. This can inform regulators and policy-makers on the lender's decision-making process and the plausible alternatives that were explored, enabling them to make concrete recommendations on what is an acceptable level of risk in meeting each objective. While this analysis was limited to five algorithms, any model, from rules-based credit score cards to neural networks to ensemble models, can be added to the trade-off curve. The process can also be adapted to decisions in other domains, such as hiring, pricing, and criminal recidivism.

The controversy of discriminatory mortgage lending predates the introduction of algorithms for credit risk evaluation, but the latter has more recently sparked a debate on how to formalise fairness as a mathematical condition for algorithms to satisfy in its predictions. A key risk of algorithms is that it may reinforce existing biases in the data, system, and society and exacerbate racial inequality. However, algorithms are not necessarily more discriminatory than existing processes. If implemented with caution, its auditability presents an opportunity for the market, lenders, and customers to consider all available options within real-world constraints and define what they want from a fair system.

Funding No funding was received for this research.

Availability of Data and Materials Data publicly available at https://www.consumerfinance.gov/data-research/hmda/historic-data/.

Compliance with Ethical Standards

Conflict of interest Michelle Seng Ah Lee is employed part-time at Deloitte LLP as a Manager in Risk Analytics. Luciano Floridi holds the following positions: Chair of the Advisory Board on Tech Ethics within APPG on Data Analytics, UK Government. Chair of the Ethics Committee of the Machine Intelligence Garage project, Digital Catapult, UK Innovation Programme. Chair of the Data Ethics Group, The Alan Turing Institute. Member of the UK ICO's Technology Advisory panel. Member of the board of the Centre for Data Ethics and Innovation. Member of the World Economic Forum's Council on the Future of Technology, Values, and Policy. Member of the Council of Europe's Expert Committee on human rights dimensions of automated data processing and different forms of artificial intelligence (MSI-AUT), Ministers' Steering Committee on Media and Information Society (CDMSI). Member responsible for the Science Panel of the Commitment to Privacy and Trust in Internet of Things security (ComPaTrioTS) Research Hub (a project funded by the British Government to work on cyber security, designing in trust, privacy, security and resilience associated with the Internet of Things).

Appendix

Features

Table of features in HMDA data, found in: https://www.ffiec.gov/hmdarawdata/FORMATS/2011HMDACodeSheet.pdf (see Table 2).

Table 2 Features

Feature	Type	Values
Income	Numeric	,000 s
Sex	Categorical	female, male, unknown, not applicable
Race	Categorical	White, black/African American, African Indian/ Alaskan Native, Asian, Native Hawaiian/ Pacific Islander, unknown
Ethnicity	Categorical	Hispanic, non-Hispanic
Agency	Categorical	Consumer Financial Protection Bureau, Dept of Housing and Urban Development, Federal deposit insurance corporation, Federal Reserve System, National Credit Union Administration, Office of the Comptroller of the currency
Owner occupancy	Categorical	Owner-occupied, not owner-occupied, not applicable
Property type	Categorical	Manufactured housing, one-to-four family dwelling
Loan purpose	Categorical	Home improvement, Home purchase, Refinancing, Conventional
Loan type	Categorical	Conventional, FHA insured, FSA RHS guaranteed, VA guaranteed
Loan amount	Numeric	,000 s
Population	Numeric	Total population in the census tract
Minority population	Numeric	%: percentage of minority population to total population for tract
FFIEC median family come ethnicity	In-numeric	USD Median family income in dollars for the MSA/MD in which the tract is located (adjusted annually by FFIEC)
Tract to MSA/MD median numeric family income percentage		% of tract median family income compared to MSA/MD median family income
Number of owner occupied numeric units		Number of dwellings, including individual condominiums, that are lived in by the owner
Number of 1- to 4-family numeric units		Dwellings that are built to house fewer than 5 families

References

Amidži, G., Massara, A. & Mialou, A. (2014). *Assessing countries' financial inclusion standing—A new composite index*. International Monetary Fund. https://books.google.co.uk/books?id=FmUcAwAAQBAJ.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev., 104*(2016), 671.

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A,, Nagar, S., et al. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018).

Berkovec, J.A., Canner, G.B., Gabriel, S.A., & Hannan, T.H. (1996). Mortgage discrimination and FHA loan performance.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Collins, P. H. (2002). *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. London: Routledge.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *u. Chi. Legal f.*, *139*.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy Non-Discrimination in Data-Driven Systems. *CoRR* arXiv preprint arXiv:1707.08120.

Demirguc-Kunt, A., & Klapper, L. (2012). *Measuring financial inclusion: The global findex database*. The World Bank.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214–226). ACM.

Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259–268). ACM.

Friedler, S.A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. arXiv preprint arXiv:1609.07236.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2017). Predictably unequal? The effects of machine learning on credit markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3072038.

Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *CoRR* abs/1710.03184. arXiv:1710.03184.

Grgic-Hlaca, N., Redmiles, E.M., Gummadi, K.P., & Weller, A. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerlan (pp. 903–912). https://doi.org/10.1145/3178876.3186138.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *CoRR* abs/1610.02413. arXiv:1610.02413.

Heidari, H., Loi, M., Gummadi, K.P., & Krause, A. (2018). A moral framework for understanding of fair ml through economic models of equality of opportunity. arXiv preprint arXiv:1809.03400.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pp. 4842–4852.

King, R. G., & Levine, R. (1993). Finance and growth: Schumpeter might be right. *The Quarterly Journal of Economics, 108*(3), 717–737.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *Aea Papers and Proceedings* (Vol. 108, pp. 22–27).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.

Koren, J.R. (2016). What does that Web search say about your credit? https://www.latimes.com/business/la-fi-zestfinance-baidu-20160715-snap-story.html.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering, 160*(2007), 3–24.

Kusner, M.J., Loftus, J.R., Russell, C., & Silva, R. (2017). Counterfactual fairness. arXiv e-prints, arXiv:1703.06856.

Ladd, H. F. (1998). Evidence on discrimination in mortgage lending. *Journal of Economic Perspectives, 12*(2), 41–62.

Lee, M. S. A. (2019). Context-conscious fairness in using machine learning to make decisions. *AI Matters, 5*(2), 23–29.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57.

Mariela, D.B. (2018). *Ethnic and Racial Disparities in Saving Behavior*. Working Papers 2018-02. Banco de México. https://ideas.repec.org/p/bdm/wpaper/2018-02.html.

Munnell, A.H., Tootell, G.M., Browne, L.E., McEneaney, J. (1996). Mortgage Lending in Boston: Interpreting HMDA data. *American Economic Review*.

Narayanan, A. (2018). Tutorial: 21 definitions of fairness. https://www.youtube.com/watch?v=jIXIuYdnyyk.

Ramadan, Q., Ahmadian, A.S., Strüber, D., Jürjens, J., & Staab, S. (2018). Model-based discrimination analysis: a position paper. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 22–28). New York: IEEE.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why should i trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13–17, 2016* (pp. 1135–1144).

Robinson, J. K. (2002). Race, gender, and familial status: Discrimination in one US mortgage lending market. *Feminist Economics, 8*(2), 63–85.

Ross, S. L., Turner, M. A., Godfrey, E., & Smith, R. R. (2008). Mortgage lending in Chicago and Los Angeles: A paired testing study of the pre-application process. *Journal of Urban Economics, 63*(3), 902–919.

Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577.

Scalia, J., Judish Baum, J., & Stute, D. (2015). Supreme court affirms FHA disparate impact claims.

Sunstein, C. R. (1994). The anticaste principle. *Michigan Law Review, 92*(8), 2410–2455.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1–7). New York: IEEE.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 2018.