AI & SOCIETY
https://doi.org/10.1007/s00146-023-01676-3

OPEN FORUM

Algorithmic discrimination in the credit domain: what do we know about it?

Ana Cristina Bicharra Garcia[1,3] · Marcio Gomes Pinto Garcia[2,3] · Roberto Rigobon[3]

Received: 8 December 2022 / Accepted: 20 March 2023

Abstract
The widespread usage of machine learning systems and econometric methods in the credit domain has transformed the decision-making process for evaluating loan applications. Automated analysis of credit applications diminishes the subjectivity of the decision-making process. On the other hand, since machine learning is based on past decisions recorded in the financial institutions' datasets, the process very often consolidates existing bias and prejudice against groups defined by race, sex, sexual orientation, and other attributes. Therefore, the interest in identifying, preventing, and mitigating algorithmic discrimination has grown exponentially in many areas, such as Computer Science, Economics, Law, and Social Science. We conducted a comprehensive systematic literature review to understand (1) the research settings, including the discrimination theory foundation, the legal framework, and the applicable fairness metric; (2) the addressed issues and solutions; and (3) the open challenges for potential future research. We explored five sources: ACM Digital Library, Google Scholar, IEEE Digital Library, Springer Link, and Scopus. Following inclusion and exclusion criteria, we selected 78 papers written in English and published between 2017 and 2022. According to the meta-analysis of this literature survey, algorithmic discrimination has been addressed mainly by looking at the CS, Law, and Economics perspectives. There has been great interest in this topic in the financial area, especially the discrimination in providing access to the mortgage market and differential treatment (different fees, number of parcels, and interest rates). Most attention has been devoted to the potential discrimination due to bias in the dataset. Researchers are still only dealing with direct discrimination, addressed by algorithmic fairness, while indirect discrimination (structural discrimination) has not received the same attention.

Keywords Algorithmic discrimination · Machine learning · Algorithmic bias · Mortgage lending · Fairness

1 Introduction

Credit is the blood of the economy. Its efficient provision is critical to economic growth and the creation of jobs. Regarding credit to families, a longtime public policy objective has been to avoid discrimination in credit provision, especially in mortgages.

Banks and financial institutions are a particular type of business because they intermediate money deposited by people or other companies. As with any business, profit is their goal. They invest the money deposited with them. Government central banks restrict their behavior to prevent too risky operations that may lead to their liquidation, affecting the entire economy. Lending money to people with a high probability of default increases the managers' liability in case of bank failure.

The approval of a loan depends on various borrowers' characteristics that reflect their ability and willingness to pay

✉ Ana Cristina Bicharra Garcia
cristina.bicharra@uniriotec.br

Marcio Gomes Pinto Garcia
mgarcia@econ.puc-rio.br

Roberto Rigobon
rigobon@mit.edu

[1] Departamento de Informática Aplicada, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro 22290, RJ, Brazil

[2] Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro 22451, RJ, Brazil

[3] Sloan School of Management, Massachussetts Institute of Technology, Cambridge, MA 02142, USA

Published online: 17 May 2023

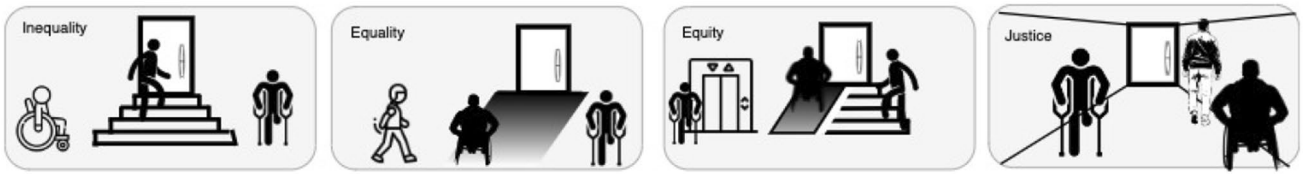


Fig. 1 Inequality vs equality vs equity vs justice: a visual explanation

the debt. One of the essential characteristics considered is the loan applicant's credit history. Unfortunately, this information is not always available. For example, immigrants, students, and young professionals take time to build a credit history. Moreover, most poor people are invisible to banks. It has been challenging for banks to deal with this lack of information. Fintechs are addressing this issue by including other types of information, such as applicants' behavior in social media and Telecom payments.

A credit analysis determines the degree of risk rating to assign to a loan applicant. Several applicants' characteristics are morally accepted by society to differentiate applicants. On the other hand, discriminating against race, sex, sexual orientation, age, human disability, religion, or marital status is a crime that must be identified and punished. For example, in 2019, Wells Fargo Bank wrote the City of Philadelphia a check for $10 million to settle a lawsuit alleging that the bank engaged in discriminatory lending practices.[1]

To empirically identify discrimination, however, is no easy task. Machine learning systems have been broadly used to suggest or decide upon loan approval. These systems learn from datasets that register lenders' decisions on past loan applications. Consequently, these systems consolidate any existing discrimination behavior under the veil of outcomes' precision and accuracy.

Fighting discrimination became a priority all over the world. For instance, in the US, in addition to laws aimed at preventing discrimination by ethnicity, gender, age, and religion, there are specific laws for the financial sector, such as the ECOA (Equal Credit Opportunity Act) and the FHA norm (Fair Housing Act), to prevent prejudice towards minorities. According to Barocas and Selbst (2016), these laws establish two legal doctrines:

- Disparate treatment: decision explicitly takes into account group membership (direct or indirectly), and
- Disparate impact: decision outcomes disproportionately hurt (or benefit) individuals of certain groups or with certain sensitive attribute values.

[1] https://www.inquirer.com/real-estate/housing/philadelphia-settles-lawsuit-wells-fargo-allegations-discriminatory-mortgage-lending-minorities-20191216.html.

The objective of this paper is to present a systematic literature review of current research dealing with identifying, preventing, and mitigating discrimination in the credit domain. We analyzed papers from 2017 to 2022. We raised two research questions:

1. What were the research settings?
   - What was the discrimination theory grounding the research?
   - What was (if any) the legal framework grounding the research?
   - What was the research perspective (domain area) when addressing the algorithmic discrimination topic?
2. Which issues were addressed?
   - What were the specific research topics addressed?
   - What was the research contribution?
3. Which open questions still need to be addressed?

We reviewed a set of 78 papers on algorithmic discrimination in the credit domain that either formalize the concept or present methods for identifying, preventing, and mitigating discrimination. Although some papers refer to justice, the authors meant equality and, at most, equity. Equality, equity, and justice are three different concepts guiding the papers. As illustrated in Fig. 1, equality implies providing people with the same resources. Equity means to provide people with the number of resources each one needs to achieve their goal. Justice refers to providing people with means so they will all have the same opportunity to achieve their goals. While equality and equity may be addressed through fair algorithms and methods, justice requires an outside agent, such as an affirmative action law. Imagine a country in which the government dedicated most of the education funds to subside public universities, aiming at providing high-quality college education. Since resources are limited, federal universities can hold only a limited number of students. Assume further that to enter these universities, one has to be well ranked in a national exam. This policy tends to increase social inequality, giving better chances to well-off individuals that went to good and expensive high

schools. An example of a policy measure to promote equality (in a partial equilibrium context) would be to increase the federal universities' enrollment so that there is a place for all students (very expensive). Examples of a strategy to promote equity could be implementing an affirmative action program, such as a quota system favoring economically challenged students or granting scholarships for poor students that get into private universities. Finally, an example of a strategy to promote justice could be to allocate funds to invest in fundamental and high school education so that all students at the college entrance level would have a similar opportunity to enter federal universities. While justice requires policymakers to act, fair algorithms may lead to equity.

The remainder of the paper is organized in the following way. Section 2 presents background knowledge to understand discrimination in the context of the credit domain, the possible sources for discrimination, and the different meanings attributed to fairness in the literature. Section 3 brings the research method followed in this systematic literature review. Section 4 presents the analysis of the overall set of papers, highlighting the research findings and the pieces of evidence to tackle the research questions. Section 5 presents a discussion of issues that have not yet been sufficiently researched, the challenges involved, and a few concluding remarks.

2 Background

2.1 The effects of discrimination in the credit domain

This section discusses the outcomes of a loan application. Disparate impact and disparate treatment play important roles.

2.1.1 Access discrimination

The most important outcome of a loan application analysis is the approval or rejection of the loan. Rejecting a loan application means an applicant cannot access the credit line service. The credit analysis is based on the assigned risk of default of applicants and the threshold of the maximum risk defined by the bank.

Applicants are allowed to question the reasons for the rejection. In general, the applicant's credit score is the most important reason. Nevertheless, the credit scoring technology is usually proprietary, with the methodology not publicly available.

The credit score is used to determine the loan concession. It is also crucial to decide on the payment conditions, such as the interest rate, the maximum number of installments, and the collateral requirement. Pricing, as well as the denial of credit, may constitute a form of discrimination.

2.1.2 Price discrimination

Price discrimination in credit markets has been chiefly associated with strategies that harm minorities or specific groups for a long time (Ladd 1998), for different types of credit, such as auto loans (Charles et al. 2008), business loans (Alesina et al. 2013), and mortgages (Bartlett et al. 2022). Nevertheless, price discrimination has been a successful selling strategy, not necessarily unethical. Considering Stigler's definition (Stigler 1987), price discrimination happens when similar goods are sold at different prices although produced at similar marginal costs. There are three types of price discrimination, depending on the amount of information available to sellers to guess the value customers assign to the product.

First-degree price discrimination, also called perfect discrimination, happens when the sellers have perfect information on how each customer values the same product for charging the maximum price. This strategy is challenging because such detailed information is difficult to obtain. And when it becomes available to sellers, it may get customers angry, especially when sellers are getting information from users without their explicit consent. In 2000, Amazon experimented with selling the same DVD with very different discounts for different users. It used data on each customer's previous purchases and navigation patterns to set the discounts (Streltfeld 2020). The strategy was discovered and made the news. Customers were angry. Amazon had to apologize and discontinue the strategy.

In the credit domain, first-degree price discrimination refers to offering different payment conditions depending on the bank's assessment of the likelihood of repayment. For example, a lower interest rate can be offered when payment is automatically deducted from the client's salary (possible in some countries).

Second-degree price discrimination refers to different prices according to different amounts of the product being sold, such as discounts for larger purchases or rewards for the next purchase. This strategy discriminates the product price, increases revenues, and is well accepted by customers.

In the credit domain, second-degree pricing discrimination refers to offering different payment conditions depending on the client's relationship with the bank (number and volume of bank products consumed).

Third-degree price discrimination refers to charging customers differently, for similar products, according to the group they belong to, inferred by attributes of the group, such as location, age, sex, and economic status. There are acceptable business examples of this type of price discrimination, such as software pricing depending on whether it will

be used for educational or professional purposes or senior discounts in theaters. Furthermore, there are examples in which price discrimination aims at better global welfare, such as the example presented in Elegido (2011) transcribed below:

> *"A young doctor in a developing country is looking for ways to establish a medical practice in the rural community where she was born, but cannot find a way to make the practice economically viable. She can see, on average, 400 patients per month. So, to cover her costs of $4000 per month (which includes her modest salary), she should charge, on average, at least $10 per visit. However, most people in her community can afford to pay, at most, $5 per visit. An economist friend suggests that she charges 90 percent of her patients only $5 per visit but charges $55 per visit to the 10 percent of her patients who can afford to pay this amount. This way, she could cover all her costs, and the rural practice would be viable. Of course, poor patients like this solution. The rich patients also like it: they would rather pay $55 per visit than travel by bad roads to the nearest hospital 50 km away, and they also like the bonus of having a doctor close at hand in case of an emergency. The doctor also is happy: this solution would allow her to practice medicine in her own community. "*

Although there are positive examples, third-degree price discrimination may harm minorities and increase structural social inequalities. In the credit domain, third-degree price discrimination refers to offering different payment conditions to clients depending on attributes that reflect the client's group, such as race, gender, age, marital status, address, and education level. The following section will address the sources of algorithmic discrimination.

2.2 Sources of algorithmic discrimination

Any software, including machine learning software, goes through a development lifecycle that starts with identifying the needs and specifying the requirements for deployment and maintenance phases. There are many different development lifecycle models, including waterfall, spiral, unified, and rapid application development, including the current in fashion agile, extreme programming, and scrum (Ruparelia 2010). Discrimination might appear in any software development phase, as illustrated in Fig. 2 and summarized in Table 1.

Discrimination can be triggered in the software specification phase as the owner and companies define goals that exclude groups on purpose. For example, many companies

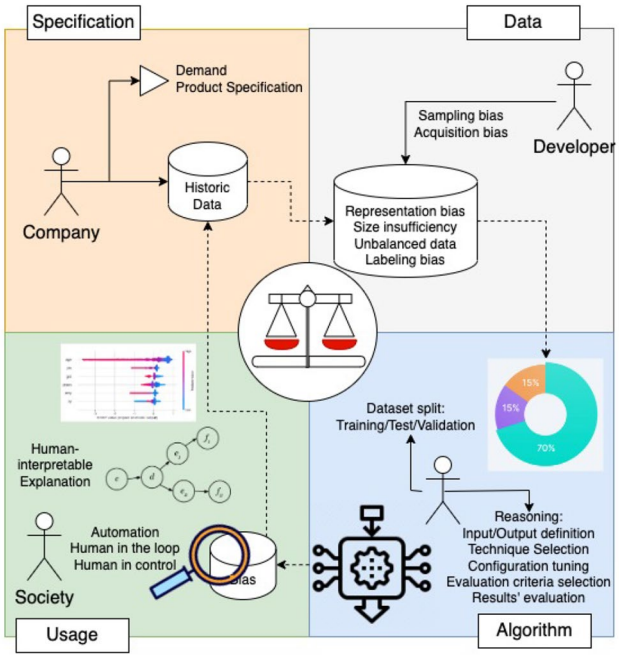


Fig. 2 Sources of algorithmic discrimination

selling games require registered users to be older than 18.[2] Another example is denying credit to people living in certain zip codes (Alliance 2014; Atkins et al. 2022). While the first kind of discrimination is acceptable, even sometimes required by law, the latter is illegal. In general, corporate decisions are registered in datasets comprising the know-how of the businesses.

Discrimination may also occur during the data processing phase, in which developers may include bias via the data acquisition and sampling method, the appropriateness of the labeling for the training data, and the dataset representativeness (Cai et al. 2020). Developers should first understand the dataset and the risks involved in using it. For example, an American health insurance company aiming at starting a preventive health care program used the number of medical appointments per year as a proxy for health fragility. Because of this procedure, it mistakenly offered preventive treatment, primarily to white patients. Blacks in that area were poorer and avoided medical visits because of the deductible and difficulties of leaving work for medical appointments. The data representation led to this misunderstanding that harmed blacks equally subjected to the considered sickness (Obermeyer et al. 2019).

The model development phase (training phase) is also a source of discrimination. The technique selection, the blindly looking for accuracy imprinted in the training, the parameter configuration, and the sampling methods to deal

[2] https://www.riotgames.com/en/terms-of-service#id.3b0qigh2mt95.

Table 1 Source of discrimination considering the system development phase

System development phase	Discrimination source	Example	References
System specification (planning)	System owners' intentional discrimination	Bank credit denial according to applicants' zip code	(Alliance 2014; Atkins et al. 2022)
Data Preparation	Poor dataset evaluation and preparation: missing data, groups under-represented, unbalanced datasets, untrustworthy labels, unidentified proxies, improper sampling methods	Bank credit denial for under-represented groups in the dataset	(Cai et al. 2020)
Model Development	Inadequate learning parameters configuration, learning technique selection, sampling methods, and learning optimization criteria (focus only on accuracy)	Bank credit denial due to decision threshold	(Schoeffer et al. 2021)
Deployment	Human actions towards automated decision or computer suggestion	Bank credit denial without looking at social impact	(Aztiria et al. 2010; Dikmen and Burns 2016; Gogoll and Müller 2017)

with unbalanced datasets are also sources of unintended discrimination (Schoeffer et al. 2021).

Last but not least, discrimination can be triggered by the usage of the computer systems' results. Decision-making can be automated, letting the computer implement its results, such as deciding to order milk as it predicts from the owners' behavior that they will run out of milk soon, (Aztiria et al. 2010), or the autonomous driving systems (Dikmen and Burns 2016). Discrimination in the latter case can be a matter of life and death since the car's image recognition might not be tuned to recognize well blacks crossing the streets (Gogoll and Müller 2017). Furthermore, results can also be presented to humans, letting them be in the decision-making loop or even be in control, auditing results and requiring explanations for a better understanding of the results. Algorithmic fairness is the subject of the next Section.

2.3 Algorithmic fairness concepts

We start this section by introducing the meaning of words that are commonly used to represent fairness: equality, equity, and justice.

According to Kassam and Marino (2021), there is no unique meaning for algorithmic fairness. Table 2 describes different metrics for fairness, reflecting the different meanings, their limitations, and the kind of discrimination they are addressing. Figure 3 illustrates the differences in the statistical meaning of the different equity concepts. In Fig. 3, we make use of the Receiver Operating Characteristic (ROC) curve for classifiers for two groups to make clearer the different equity concepts. The ROC curve was built considering the classifier's performance for classifying individuals of a group as belonging to a class or not, such as a class of "good payers" in the credit domain. The confusion matrix is the basis for building the ROC curve. A true positive means the loan was approved for a creditworthy person; a false positive means the loan was approved for a defaulter. The area under the ROC curve (AUC) reflects the quality of the classifier, a purely random classifier would produce a diagonal straight line. We used the ROC curves to exemplify the different notions of equity.

Let us examine the statistical meaning of the several fairness criteria portrayed in Fig. 3. The Equality of Opportunity fairness criterion is akin to requiring that the classifier for both groups produce the same true positive rates. This means, for example, in the case of gender discrimination, that, irrespective of gender, a creditworthy person would have the same probability of getting the loan application approved. Note, however, that, in general, the false positive rates would differ, i.e., the bank would expect more defaults from one group than from the other. Equality of Odds, on the other hand, requires that both the true positive and the false positive rates be equalized. This is only possible if the two ROC curves intercept, which may not always happen. Even if they intercept, the resulting true and false positive rates may not be compatible with the bank's economic objectives. For example, assume that both ROC curves intercept at the 25 percent true positive rate. This true positive rate would probably be too low for any reasonable credit scoring system. The Demographic Parity fairness concept assumes that the bank sets up a maximum false positive rate, and applies such rate to both groups. Assuming that the creditworthiness of both groups is different, such a procedure implies possibly different credit score thresholds for the acceptance of loan applications from the two groups. People from the more creditworthy group would have their loan applications denied, while people with similar credit scores from the less creditworthy group would have their applications accepted.

Table 2 Fairness definitions and limitations

Name	Description	Limitation	Obs
Veil of ignorance (Grgic-Hlaca et al. 2016; Ladd 1998)	Remove protected attributes; functional interdependence premise	Proxy discrimination (Datta et al. 2017)	Individual-oriented ("disparate impact")
Equalized Odds (Hardt et al. 2016)	Equal true positive rates and equal false positive rates across groups	It may ratify structural prejudices	Group-oriented
Demographic Parity - Statistical parity (Feldman et al. 2015)	Distribution of outcomes mirrors population distribution	Accuracy decay; performance instability	Group-oriented (US Equal employment opportunity Act 1978)
Calibration (Chouldechova 2017; Liu et al. 2017; Corbett-Davies and Goel 2018)	Equal predicted credit score probability leads to equal probability of "default" (for binary classifiers)	Calibration is open to manipulation of the risk distribution for different groups	Group-oriented
Equality of Opportunity (Hardt et al. 2016)	Equal true positive rates across groups	It may ratify structural prejudices	Group-oriented Disability-based equality of opportunity
Equality of Treatment (Berk et al. 2021)	Similar ratio false negative and false positive rates across groups	Accuracy decay	Group-oriented
Counterfactual fairness (Russell et al. 2017)	Causal model of the world: Maintenance of outcomes in face of changes of values in the protected attributes	Identifying the attributes and proper counterfactual model	Individual-oriented "disparate impact"
Individual fairness (Dwork et al. 2012)	Similar individuals get similar outcomes	Finding proper metric for similarity between individuals	Individual-oriented
Predicted parity (Zafar et al. 2017; Chouldechova 2017)	Same precision rate across group	It may ratify structural prejudices	Group-oriented
Meritocracy Fairness (Kearns et al. 2017; Kearns 2017; Joseph et al. 2016)	If quality of A is at least as good as quality of B, then choose A with at least the same probability as choosing B	It may disguise prejudices	Group-oriented

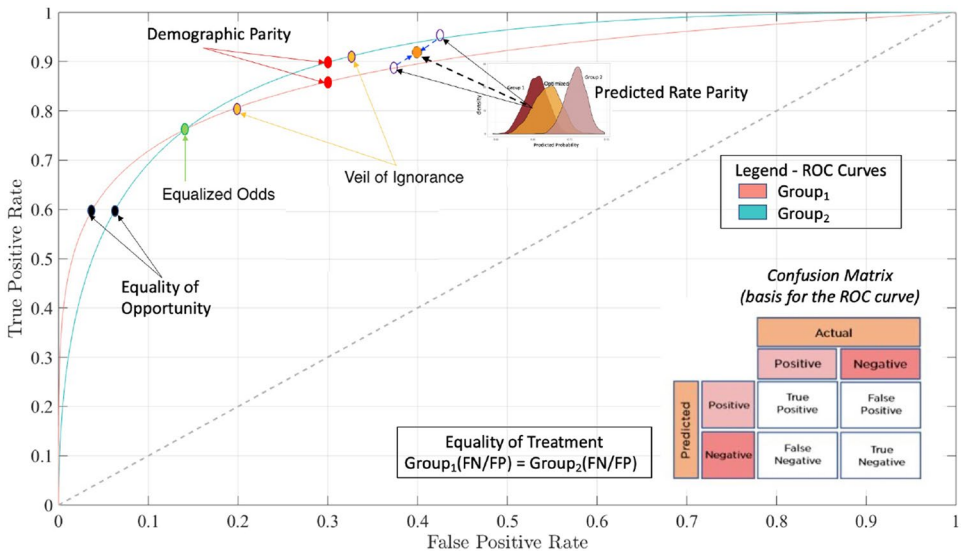


Fig. 3 Different notions of equity explained via ROC curve behavior

And the true positive rates would also be presumably different unless this criterion coincided with the Equality of Odds criterion. The Equality of Treatment criterion requires similar ratios of false negative to false positive rates across groups.

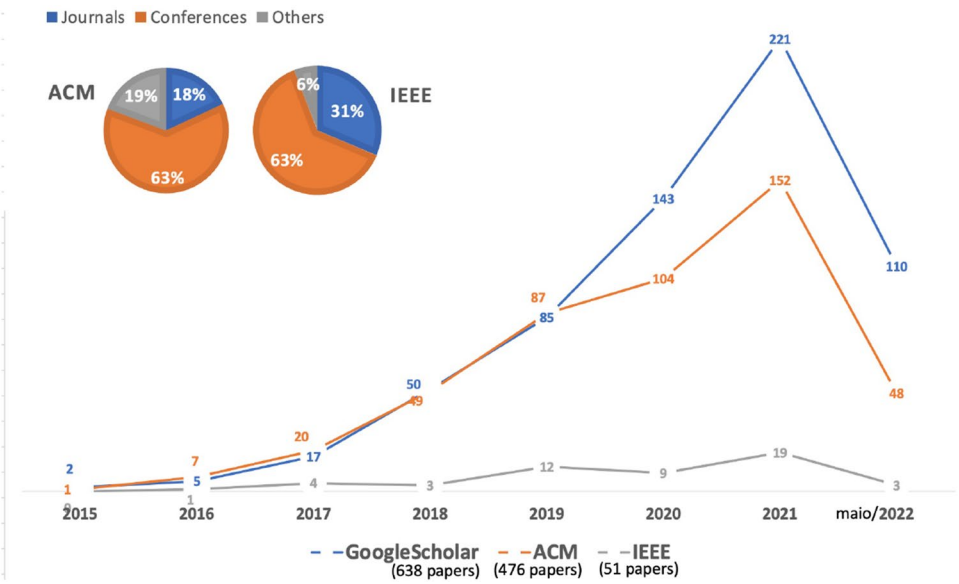


Fig. 4 Number of publication addressing "algorithmic discrimination" or "algorithmic bias" through the years

It is almost impossible[3] to accomplish the requirements for all definitions (Chouldechova 2017; Pleiss et al. 2017). Trade-offs between fairness and accuracy (Corbett-Davies and Goel 2018) must be considered in each application domain, from loans to job applications and parole decisions. No matter how many definitions of fairness we can arrive at, they will remain contestable vis-à-vis some other definitions.

Programming fair algorithms hinges on the definition of fairness, which varies, as shown in Table 2. Algorithmic fairness can be accomplished through strategies during the data collection, pre-processing, at-processing, or/and post-processing treatments. During data collection, the focus should be on guaranteeing or verifying the representativeness of the data. Are the instances (cases) properly described using all needed variables? Are the set of instances representative of the population? These are usual issues for any robust statistical analysis. Whenever data collection is not adequate, imperfect labeling is an issue. Strategies to deal with biased labels include verifying the compatibility of the outcome with other features to deal with meritocracy unfairness (Schoeffer et al. 2021; Wang and Gupta 2020). Pre-processing strategies for algorithmic fairness refer to dataset manipulation, such as changing labels in randomized data points (Calmon et al. 2017; Gordaliza et al. 2019), adding synthetic minority class examples ("over-sampling technique") (Chawla et al. 2002; Chakraborty et al. 2021), removing examples of majority class ("under-sampling technique") (Elhassan and Aljurf 2016), re-weighting data pairs (Kamiran and Calders 2012), or combination of techniques (Banasik and Crook 2007). At-processing approaches refer to changing the way the computation learning process works, such as by including a social welfare constraint (Cohen et al. 2022) or a regularization term to the existing optimization objective function (Zafar et al. 2017), using adversarial debiasing (Zhang et al. 2018) or even analysis of counterfactual (Russell et al. 2017). Finally, post-processing approaches refer to changes in the decision function after the training process, such as finding a proper threshold that would allow a fairer result (Hardt et al. 2016) or trading off accuracy and fairness (Chen et al. 2018).

3 Systematic review method

This research followed the Kitchenham et al. systematic literature review protocol (Kitchenham et al. 2009) for planning, conducting, and reporting the results. This section describes the main activities in the review planning and conducting phase.

3.1 Planning phase

The planning phase encompasses the identification of the research questions, the search strategy, the selection of the proper search keywords, and the definition of the inclusion and exclusion criteria. Since we are interested in algorithmic discrimination, we first apply a more generic search on Google Scholar, ACM, and IEEE Digital libraries to check the search space and the research interest evolution through the years. Figure 4 illustrates the increasing interest in "algorithmic discrimination" or "algorithmic bias" topics. As shown below, we better delineate the search scope to address our research objective.

[3] Only in perfect conditions of a perfect predictor. (Miconi 2017).

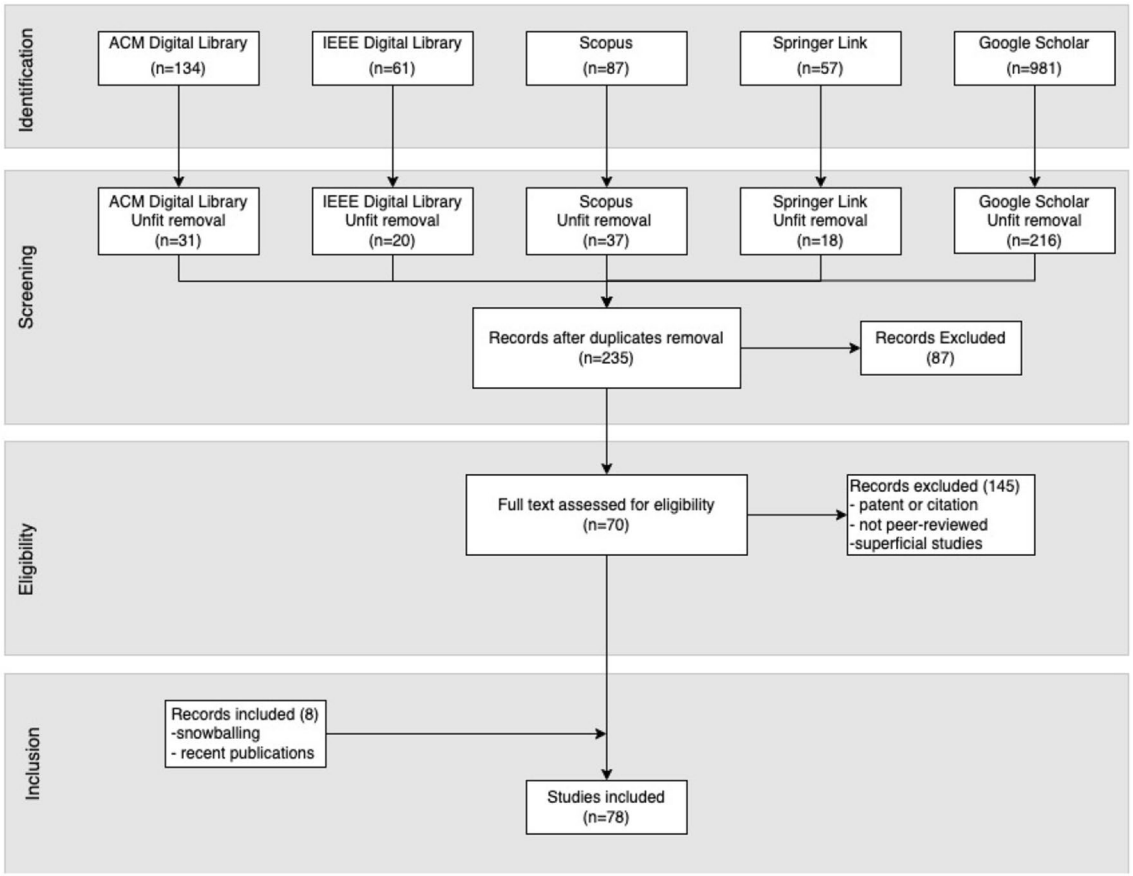


Fig. 5 Literature review process

To address our research questions, we used the search keywords were:

- "algorithmic discrimination" and
- "bank loan".

The initial search identified related expressions, such as: "algorithmic fairness" and "discriminatory" for "algorithmic discrimination" and "mortgage" for "bank loan". Additionally, "loan" was more general than "bank loan" and was used instead. These changes helped tune the formulated search string to:

(("loan") OR ("credit") OR ("mortgage")) AND
(("algorithmic discrimination") OR ("algorithmic fairness") OR ("discriminatory"))

The search string was adjusted to fit the database's search format. Papers were retrieved from the following databases: ACM Digital Library, Google Scholar, IEEE Digital Library, Springer Link, and Scopus.

3.2 Conducting phase

After conducting an automatic search of the databases, the three authors reviewed the retrieved papers to verify the paper's conformance to the inclusion and exclusion criteria, previously agreed upon. The paper selection process follows the strategy presented in Fig. 5. *Parsifal*[4] and *Publish or Perish*[5] tools were used to automatically search the papers.

The inclusion criteria consisted of the following rules:

- I1: describes a model for defining, detecting, preventing, or mitigating discrimination in credit domains,
- I2: describes parameters that explain discrimination in the bank credit domain,
- I3: is published as a conference paper, journal article, book chapter, or technical report,
- I4: is published between 2017 and 2022,

[4] https://parsif.al/.

[5] https://harzing.com/resources/publish-or-perish.

Additional papers were included as a result of finding relevant material from citations in the original list of retrieved papers, in a process called snowballing (Wohlin 2014).

The set of exclusion criteria are:

- E1: describes discrimination only theoretically,
- E2: does not present empirical data,
- E3: focuses on system implementation without highlighting the discrimination aspect,
- E4: describes a general proposal without describing a discrimination model or implementation details,
- E5: is published as a complete book, presentation slides, editorial, thesis, or has not been published yet,
- E6: is not written in English,
- E7: cites credit loans, but is not applicable to the domain.

The initial selection contained 1320 papers. After applying the exclusion criteria, we reached a set of 78 papers, including journal papers and conference papers from 2017 to April of 2022. Three studies were included outside this period for being relevant to this review.

4 Findings

This section presents the analysis and the findings for answering the two research questions guiding this literature review. Appendix A presents the details of each of these papers.

4.1 Preliminary analysis

According to the publication source and declared objective, the selected papers were classified into five domain areas: computer science, economics, law, operational research, and philosophy. As shown in Table 3, most papers brought computer science or economics perspectives.

We merged the abstracts of all papers, removed the stop words, and created a word cloud. Figure 6 presents the most frequent words that appear in the set of papers. The words are not comprehensive but highlight the most frequent topics. A possible reading from the figure may say:

> *"The set of papers addresses the issues of fairness and discrimination coming from data and algorithms that impact decisions causing consequences. Discrimination is mostly racial and gender based. Credit loans and mortgages are the tasks being studied in which subjects are individuals, borrowers, groups, and people. The research goals include either classification, prediction, definition, or perception of discrimination. Papers are proposing models, methods, approaches, frameworks, systems, or literature reviews. The papers talk about metrics, evaluation, accuracy, error, and performance of their proposals. Many studies are domain-, case- or application-oriented. Researchers also address the what, where, when, and how discrimination occurs."*

In general, papers are presenting research on: algorithm discrimination or algorithm fairness in which data play a very important role and impact decision-making. Mostly, discrimination is related to race and gender. Research goals have been mostly classification, but prediction and definition are also frequent. The studies mostly propose models, methods, approaches, frameworks, and systems. The subjects have been individuals, people, or groups. The task involves getting credit either a loan or a mortgage. There are case studies for specific datasets and countries. The intervention that may be the discrimination cause or the tool to find discrimination mostly involves machine learning, descriptive and inferential statistics models, and ontological models (mainly for defining fairness and discrimination).

Most countries included in our survey present case studies using inferential statistics to identify discrimination on public or private loan datasets. As shown in Fig. 7, the United States leads the research on discrimination, especially in terms of creating a theoretical framework for defining fairness metrics. European countries focus on general discrimination. Developing and low-income countries have focused on identifying discrimination, mostly sex discrimination from bank datasets. Somewhat surprisingly, the first authors of the selected papers were mostly white men, as shown in Table 4. The identification of sex and race/ethnicity of the first author of each paper is an approximation. This information was inferred by visual perception of the author's name and facial image available on the Internet.

4.2 The research settings

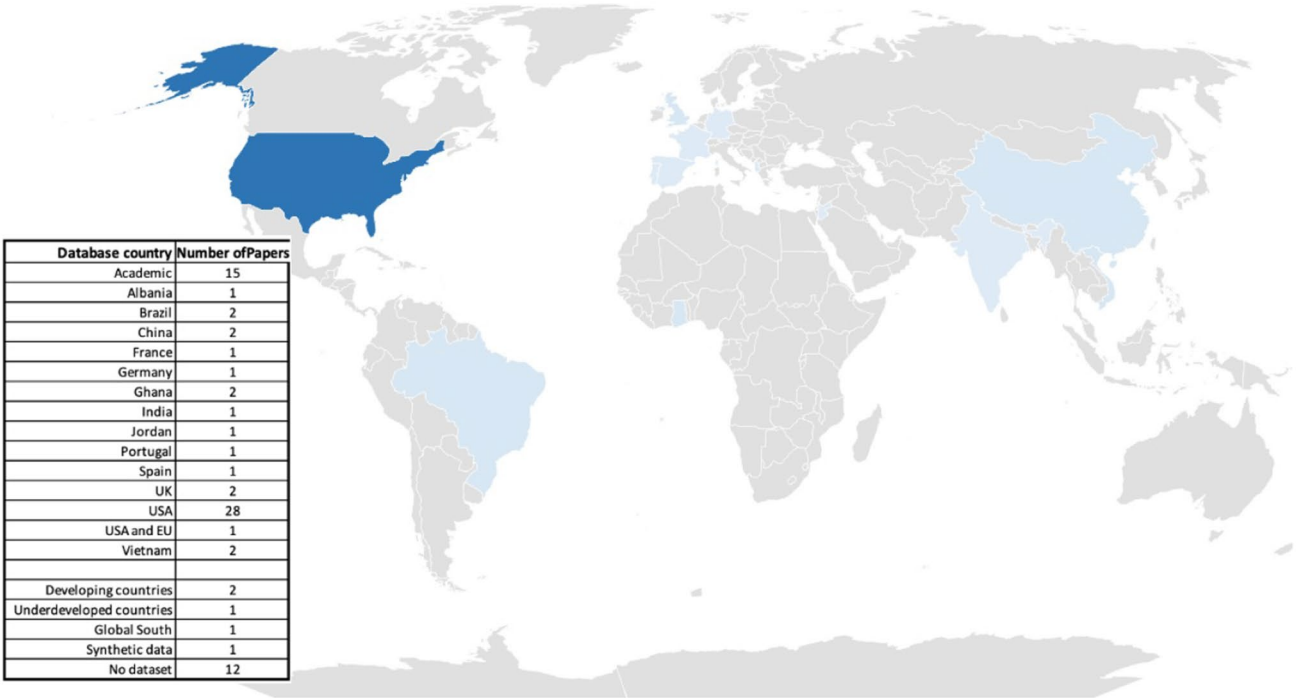
To answer our first research question concerning the research settings, we analyzed each of the 78 papers looking for descriptors for clustering or distinguishing the approaches presented in the papers. We came up with an ontology for discrimination research in the credit domain. As described in Fig. 8, research on discrimination in the credit domain has its findings delimited to a domain area and to a country from which the data came. It is developed according to a type of research, using specific research methods applied over a dataset containing data that will support the findings. The dataset records instances of loans and loan applications from a country. The data discrimination analysis is done considering fairness metrics. The research should be grounded on a discrimination theory and related to a legal framework. The research focuses on a topic delimited by the scope and looks

Table 3 Papers timeline: 2017–2022. 1999, 2003, and 2016 papers were included by snowballing (Wohlin 2014)

Area	1999*	2003*	2016*	2017	2018	2019	2020	2021	2022
CS	(Black 1999)		(Kleinberg et al. 2016)	(Kearns 2017; Chen et al. 2017)	(Binns et al. 2018; Farnadi et al. 2018)	(Elzayn et al. 2019; Saxena et al. 2019)	(Cai et al. 2020; Bogen et al. 2020)	(Karimi et al. 2021; Ghosh et al. 2021)	(Cohen et al. 2022; Singh et al. 2022)
					(Wakchaure and Sane 2018; Kallus and Zhou 2018)	(Liu et al. 2019; Valentim et al. 2019)	(Coenen et al. 2020; Ragnedda 2020)	(Schoeffer et al. 2021; Chakraborty et al. 2021)	(Kallus et al. 2022; Kordzadeh and Ghasemaghaei 2022)
						(Lohia et al. 2019; Salimi et al. 2019)		(Corrales-Barquero et al. 2021; Hassani 2021)	
						(Bellamy et al. 2019; Bryant et al. 2019)		(Lee and Floridi 2021; Albach and Wright 2021)	
						(Sun and Gao 2019)		(Hort and Sarro 2021) (Mehrabi et al. 2021)	
								(Moscato et al. 2021; Segal et al. 2021)	
ECO				(Bayer et al. 2017; Aitken 2017)	(Tran et al. 2018; Salgado and Aires 2018)	(Fuster et al. 2019)	(Sackey and Amponsah 2020; Otieno et al. 2020)	(De Andrés et al. 2021; Ambrose et al. 2021)	(Fuster et al. 2022; Loya 2022)
					(Cozarenco and Szafarz 2018; Maaitah 2018)		(Pi et al. 2020)	(Li 2021; Blanco-Oliver et al. 2021)	(Park 2022; Yu 2022)
					(Steil et al. 2018; Faber 2018)			(Mitchell et al. 2021; Bono et al. 2021)	(Brotcke 2022; Nyarko 2022)
					(Beck et al. 2018; Sackey and Amponsah 2018)			(Rebitschek et al. 2021; Giacoletti et al. 2021)	
					(Le and Stefańczyk 2018)			(Bhutta and Hizmo 2021)	
LAW					(Bruckner 2018; Cofone 2018)	(Allen 2019; Knight 2019)			(Mendes and Mattiuzzo 2022)
						(Dillbary and Edwards 2019) (Swan 2019)			
OR							(Wong 2020)		(Kozodoi et al. 2022)
PHY		(Baesens et al. 2003)							(Prasad 2022)
Total	1	1	1	4	15	14	8	21	13

Fig. 6 Word cloud containing the most frequent words that appear in the abstract of the 87 selected papers

2. the type of research: argumentative/essay, explanatory (ex-post facto), theoretical, and empirical. Depending on the type of research, one or more research methods were applied, such as literature review, essay, inferential statistics, descriptive statistics, content analysis, counterfactual analysis, and machine learning. Depending on the type of work, inferences are driven from the analysis of **datasets** that can be public, private, created by the experiments (own dataset), academic or synthetic. The academic datasets are sample datasets, properly anonymized, donated by commercial banks, and available in public repositories, such as the UC Irvine machine learning repository.[6] Two datasets are primarily used in the credit domain: the Australian dataset containing 600 instances described by 14 attributes and the German dataset containing 1000 instances described by 20 attributes. There is a benchmarking study comparing credit scoring methods (Baesens et al. 2003; Lessmann



Database country	Number ofPapers
Academic	15
Albania	1
Brazil	2
China	2
France	1
Germany	1
Ghana	2
India	1
Jordan	1
Portugal	1
Spain	1
UK	2
USA	28
USA and EU	1
Vietnam	2
Developing countries	2
Underdeveloped countries	1
Global South	1
Synthetic data	1
No dataset	12

Fig. 7 Geographic distribution of the publications considering the country of the dataset source. The "Non-specified Countries" means the paper did not use datasets either for using synthetic or academic datasets or for a theoretical argumentation

at the impact of discrimination on the lives of individuals or society.

Seven attributes describe the research settings:

1. the domain area: research published in forums such as computer science (CS), economics/business (ECO), operation research (OR), Law, philosophy, or social science (PHY).

et al. 2015) that, in addition to using the German and the Australian datasets, also used other academic datasets provided by Benelux and UK financial institutions: the Bene1 dataset containing 3123 instances described by 27 attributes, the Bene2 dataset containing 7190 instances

[6] http://archive.ics.uci.edu/ml/.

Table 4 Sex and race/ethnicity of the first authors of the retrieved papers

Total number of papers = 78		
Sex	Quantity	Percentage
Female	18	23%
Male	60	77%
Race/ethnicity	Quantity	Percentage
Asian	12	15.4%
Black	7	9.0%
Indian	14	17.9%
Latino	5	6.4%
White	40	51.3%

described by 37 attributes provided by companies for PAKDD data mining competition.[7] The datasets depict decisions on a loan from a Country.

3. the fairness metric used to evaluate the work: demographic parity, equalized odds, counterfactual fairness, predicted parity, individual fairness, and performance accuracy.
4. the discrimination theory grounding the research, shown in Table 5: taste-based discrimination (Becker 2010), information asymmetry (Stiglitz and Weiss 1992), statistic discrimination (focus on the outcome) (Phelps 1972), statistic discrimination (focus on the skills) (Arrow 2015), implicit discrimination (Bertrand et al. 2005), intersectionality discrimination (Crenshaw 1989), symbolic capital discrimination (Bourdieu 2018), visual legibility (Latour 1986), statistic sampling (Heck-

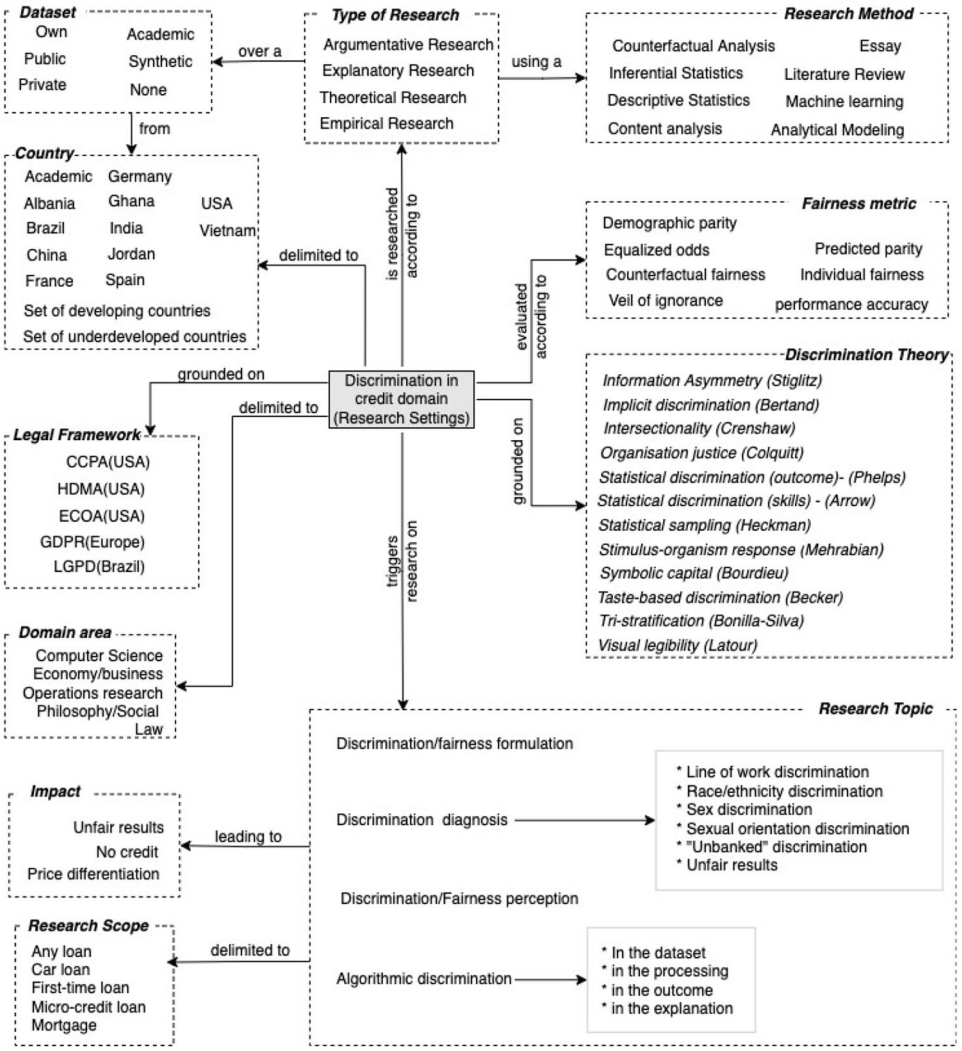


Fig. 8 An ontology for describing research on the discrimination topic

described by 18 attributes, and the UK dataset containing 30,000 instances described by 14 attributes. They also used the Pak dataset containing 50,000 instances

[7] http://sede.neurotech.com.br/PAKDD2010/.

Table 5 Theories of discrimination mentioned by the retrieved papers

Theory	Description	Author
Profit maximization (required skills)	Statistical model for profit maximization—focus on skills (e.g., current job-monthly income)	Arrow (2015)
Taste-based	People are willing to pay a premium to maintain their prejudicial preferences	Becker (2010)
Implicit discrimination	Similar to taste-based, but without intention	Bertrand et al. (2005)
Tri-racial stratification	Instead of white and black stratification white, honorary white and black (e.g., light-skinned Latinos)	Bonilla-Silva (2004)
Symbolic capital	People assign value to decisions that maintain the structure of classes whiteness as symbolic capital	Bourdieu (2018)
Intersectionality	Sex, race, social class, and sexuality can not be analyzed separately	Crenshaw (1991)
Organization justice	It is not acceptable an algorithm disproportionate benefit a group	Colquitt and Rodell (2015)
FICO	It is a personal creditworthiness index largely accepted by industry varies from 300 to 900 (the formula is confidential)	FICO@score (2022)
Statistical Model of sample adjustments	The bias that arises when using least squares in datasets with missing data	Heckman (1979)
Stimulus-organism-response	Environmental stimulus influence individual internal state that leads to behavioral responses	Mehrabian and Russell (1974)
Visual legibility	Visual inscriptions are traces of subjects' behavior (e.g., credit score for creditworthiness)	Latour (1986)
Profit maximization	Statistical model for (final outcome) profit maximization—focus on the outcome (e.g., default history)	Phelps (1972)
Information asymmetry	Transactions' participants hold distinct amounts of information of each other (e.g., lenders publicized information, borrowers without a credit history)	Stiglitz and Weiss (1992)

man 1979), race tri-stratification (Bonilla-Silva 2004). Becker's taste-based discrimination is the fundamental grounding theory pervading all discrimination research.

5. the legal framework: the USA Home Mortgage Conflict of interest Act (HMDA),[8] the USA Equal Credit Opportunity Act (ECOA),[9] the USA California Consumer Privacy Act of 2020 (CCPA),[10] The European General Data Protection Regulation (GDPR),[11] the Brazilian General Data Protection Law (LGPD).[12]
6. the research topic: formulation of fairness/discrimination concept, perception of fairness/discrimination on outcomes and decision-making process, diagnosing discrimination on datasets, and algorithmic discrimination. The research topic focuses on discrimination leading to **impacts** of either unfair results, credit rejection, or price differentiation. The research topic is delimited to a **research scope**, such as any loan, car loan, first-time loan, micro-credit loan, or mortgage. There was no educational loan mentioned in our set of reviewed papers.
7. the Country of the dataset: Albania, Brazil, China, France, Ghana, India, Jordan, Spain, the USA, a set of developing countries, loan data from an international funding agency with data from 52 developing countries, and data from micro-credit platform (Kiva) containing data from low-income countries. Thus, the academic datasets are sample datasets donated by companies. In addition, the academic datasets are from Australia, Benelux, German, the UK, the USA, and KDD competitions.

As previously observed and shown in the table 3, most papers were from computer science and economics domains, looking at generic discrimination or unfair outcomes, and specific bias concerning sex, race/ethnicity, and a few regarding sexual orientation, poor people, people without credit history and small farmers. There is no evidence of studies relating to another type of discrimination, such as religion, in credit markets.

Another interesting finding is the worldwide concern with discrimination, but with a total American dominance, in identifying discrimination by analyzing the datasets. Few studies explicitly cite the discrimination theory grounding

[8] https://www.consumerfinance.gov/data-research/hmda/.

[9] https://www.federalreserve.gov/boarddocs/supmanual/cch/fair_lend_reg_b.pdf.

[10] https://oag.ca.gov/privacy/ccpa.

[11] https://gdpr-info.eu/.

[12] http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

their work, but they were implicitly considering Becker's taste-based discrimination behavior.

4.3 The research issues

After considering the studies, we mapped the research issues into one of seven categories, as presented in Fig. 9: literature reviews, data analytic/econometric, human perception, definition, dataset, algorithmic and outcome issues. Some studies focus on more than one issue category.

4.4 Literature reviews and essays

Studies on **race and ethnicity** discrimination in credit loans are not new. Tables 6 and 7 summarize the findings in previous literature reviews and essays. In 1999, Black Black (1999) presented a **literature review** on race and ethnicity discrimination in the credit market domain. He claimed discrimination is more against properties (redlining) than individuals. He highlighted the effect of job instability and the lack of credit history as the significant factors for the loan acceptance gap between black and white applicants. Nevertheless, he did not investigate the root cause for these effects that might be explained by structural discrimination fostered by the credit approval distinction (loan rejections for blacks and Hispanics are three-to-one above for whites—**demographic parity fairness metric**).

Black Black (1999) did not consider the impact of the new techniques to calculate the credit scores of individuals that may be the cause for discrimination. Lessmann et al. (2015) reviewed new techniques, including machine learning, to assess credit scoring. They analyzed 41 different classifiers, considering the number of datasets used for testing, the number of variables per dataset, and the technique itself, such as: Artificial neural network, Support vector machine, and Ensemble classifier. They evaluated the classifiers using eight academic datasets, including the Australian credit (AC) and German credit (GC) from the UCI Library. Their results indicated that the new techniques perform better than traditional logistic regression, especially heterogeneous ensemble classifiers. Besides, they present evidence of increased financial returns with more accurate scorecards. Moscato et al. (2021) proposed a benchmark study that evaluates the performance (e.g., AUC, Sensitivity, Specificity) of different classifiers (e.g., random forest, logistic regression, and artificial neural networks) under different sampling strategies (e.g., under-sampling and over-sampling strategies) to deal with unbalanced datasets. They also considered the fit of different explicability methods, especially Lime and Shap, to offer transparency of computer decision-making. They used a public dataset from the "Lending Club" fintech marketplace bank (data available at Kaggle repositories[13]) containing 877,956 samples of loan applications.

In 2017, Aitken (2017) had already realized the usefulness of using informal information from social media as an alternative credit scoring for the *"unbanked"*. Unbanked are people without formal documentation proving their credit status, **people without credit history**, having **no access to loans**. They are considered too risky for lenders to loan them money. Based on Latour (1986) **visual legibility theory** for making visible the different perspectives of a subject, Aitken reflects upon the advantage and disadvantages of current approaches to making visible the financially invisible unbaked, such as the experiment using social behavior information—360-degree views of borrowers being experienced by FICO.[14] He brought up the paradox the literature poses in which the under-development alternative scores may work for increasing loan inclusion, in the **equity**, or ratifying the credit-worthless unbanked.

More recently, in 2021, there are reviews looking specifically at bias in credit scoring techniques. Corrales-Barquero et al. (2021) recently (2021) published a literature review on sex bias in credit scoring methods. They reviewed a set

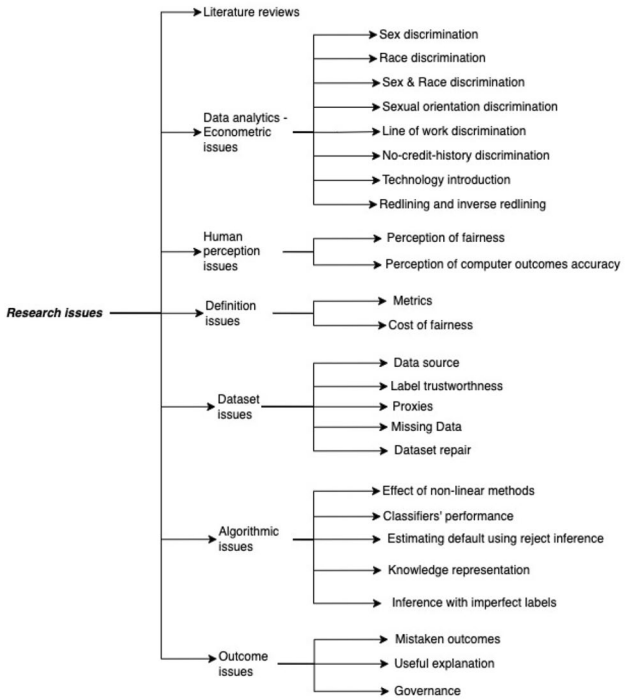


Fig. 9 Research issues discussed in the papers

[13] https://www.kaggle.com/datasets/wordsforthewise/lending-club.

[14] FICO—a company that develops the standard credit score largely used at the financial industry.

Table 6 Reviews and essays concerning algorithmic discrimination—Part 1

Refs	Scope	Research Issue	Findings
Black (1999)	Bank loans (mortgage)	Race and ethnicity discrimination	Loan rejections for blacks and Hispanics are three-to-one above whites'. Discrimination is more against properties (redlining) than individuals. Job instability and lack of credit history are the main factors for discrimination
Lessmann et al. (2015)	Bank loans	Performance of loan application classifiers	New ML techniques perform better than traditional logistic regression, especially heterogeneous ensemble classifiers
Moscato et al. (2021)	Online source (lending market)	Performance of sampling techniques with credit risk models to predict loan repayment	Under-sampling with random forest results presented the best performance (AUC) using the Lending Club public dataset
Aitken (2017)	Bank loans	Cold start: Credit score calculation for people without a credit history	Social media information can meet the banks' needs to infer loan applicants' ability and willingness to pay; i.e., augment credit scoring calculation for the "unbanked"
Corrales-Barquero et al. (2021)	Bank loans	Techniques to reduce sex bias in credit scoring models	Bias classification: disparate treatment bias (protected attributes in the dataset); associated bias (bias through proxies); selection bias (groups misrepresented in datasets), and intentional bias (purposeful prejudice). There are mitigation techniques for each type of bias
Kordzadeh and Ghasemaghaei (2022)	Algorithmic bias	Algorithmic bias via the stimulus-organism-response and organizational justice theories	Conceptual model for algorithmic bias considering fairness' perception, acceptance, and action of decision-makers towards the computer's suggestions. The model includes (1) the decision-makers' beliefs and moral identity, (2) the task automation degree, (3) the computer reasoning transparency, (4) the ethical norms of the organization, and (5) the laws and social norms of the environment
Mehrabi et al. (2021)	Algorithmic bias	Fairness for ML systems during the pre-processing, the ML training, and the post-processing phases	Fairness classification: individual fairness, including veil of ignorance and counterfactual fairness; group fairness, including demographic fairness and equalized odds; and subgroup fairness

Table 7 Reviews and essays concerning algorithmic discrimination (part 2)

Ref	Scope	Research Issue	Findings
Wong (2020)	Algorithmic bias	Techniques to remove algorithmic bias	**Techniques to remove bias** during dataset pre-processing, the machine learning phase, and the post-processing phase, without a clear definition of what bias is
Ragnedda (2020)	Algorithmic bias	Social inequalities raised by automated ML systems for tasks such as loan approval	**Inequalities source classification:** (a) Lack of unified understanding of the impact of automated decisions' outcomes for individuals and society, (b) Implicit bias imprinted in the dataset and (c) Decision-makers discriminatory behavior nudged by suggestions of intelligent systems
Mitchell and Shadlen (2017)	Algorithmic bias	Fairness for ML systems	**Fairness classification** emphasizing the implicit assumption of automated decision-making: (a) driven by a utility maximization function with a single threshold on the predictions dividing the population, (b) driven by an equal prediction measure with similar prediction impact across groups and (c) driven by an equal measure over the decisions
Bruckner (2018)	Algorithmic bias	Benefits and challenges of algorithmic lender systems	**Benefits;** faster, cheaper, more predictable credit score analysis and able include people without a credit history than conventional lenders. **Disadvantages:** consolidate decision-making patterns from datasets, perpetuating discrimination. Need specific regulation for algorithmic lenders, and adjust current American legislation, specifically the American ECOA

of 20 papers to identify the techniques that have been used to reduce sex bias in credit scoring models. They differentiated the bias into disparate treatment (when the protected attributes were present in the dataset), associated bias (due to proxies), selection bias (dataset with misrepresented groups), and intentional bias. The authors also listed mitigation techniques and the requirements for applying them.

Kordzadeh and Ghasemaghaei (2022) reviewed 56 papers on algorithmic bias through the stimulus–organism–response theory and organizational justice theory. Their objective was to understand the impact on decision-making. Based on the literature, they present a conceptual model showing algorithmic bias impacts decision-makers' fairness perception of the computational outcome and, consequently, their acceptance and adoption of the suggestions. This stimulus–response flow is influenced by individual characteristics in terms of beliefs and moral identity, task characteristics such as automated or human control, technology characteristics such as reasoning transparency, organizational characteristics, such as organizations' norms and rules, and environmental characteristics, such as laws and social norms.

Proposing solutions to deal with discrimination requires a good definition of what is the problem. For that matter, understanding what is a fair result, a fair algorithm, and a fair decision is fundamental to leading to a solution. Mehrabi et al. (2021) review the literature on fairness definition for machine learning systems and proposed a taxonomy in which fairness is divided into three groups: individual fairness, including veil of ignorance and counterfactual fairness; group fairness, including demographic fairness and equalized odds; and subgroup fairness. They reviewed different approaches for fair machine learning systems either in the pre-processing, in-processing, and post-processing phases and tested with academic datasets (UCI datasets) from a variety of domains.

Wong (2020) reviewed the **literature** on algorithmic discrimination looking at the way researchers have been addressing the problem. He distinguished the different approaches for removing bias as pre-processing, during the learning processing, or as a post-processing technique. He claimed that many technical solutions have been proposed to remove bias without a clear definition of a fairness system.

Ragnedda (2020) discussed the challenges imposed by the introduction in the society of new digital technologies, such as machine learning, to increase social inequalities. He classifies the **inequalities in three types**: (a) knowledge inequalities reflecting the different understanding among people of the impact in our lives from the outcome of these

technologies, (b) inequalities imprinted in the data guiding the learning process of systems and (c) inequalities of the treatment nudged by intelligent systems according to socio-demographic characteristics of individuals in the society. This **essay based on the literature** organizes the topic.

Mitchell et al. (2021) discussed fairness in the context of decision-making either human or computational. The authors classified the existing definition of fairness into three categories: driven by a utility maximization with a single threshold over the predictions dividing the population, driven by an equal prediction measure with similar prediction impact across groups, and driven by an equal measure over the decision. They used this classification to shed light on the relation between the implicit assumptions and choices of prediction-based decision-making and the fairness of the outcomes.

Bruckner (2018) discusses the benefits and challenges of algorithmic lender 2.0. The benefits include offering loans faster, cheaper, and with more predictive credit than conventional lenders. Algorithmic lenders broaden the range of borrowers, extending to the credit invisible for being able to gather and process loan applicants' information from varied sources including social media. Algorithmic lenders eliminate human decision-making subjectivity and discriminatory behavior. On the other hand, since these systems learn decision-making patterns from datasets, they can perpetuate discrimination. The authors emphasized the need for regulation of algorithmic lenders. They also discussed the adequacy of current American legislation, specifically the American Equal Credit Opportunity Act (ECOA). They provided examples of the two ways ECOA violation (disparate treatment and disparate impact) can be proven concerning the disparate treatment and disparate impact. For example, giving a more favorable credit term for older people is an example of disparate treatment. Defining a minimum value for a credit loan that heavily excludes a racial group, even if not intentionally discriminatory, is an example of disparate impact. Lenders can avoid liability if they can prove the discrimination has a valid purpose, other than maximizing profit. Lenders must show the relation between creditworthiness and the policy being violated.

4.5 Data analytics: econometric issues

A great deal of research focused on making sense of data and identifying trends that indicate discrimination in the decision-making process. The goal is to identify systematic discriminatory behaviors engraved in the datasets. Thirty-two of the seventy-eight reviewed papers addressed this issue, as summarized in Tables 8 and 9.[15]

4.5.1 Sex discrimination

Beck et al. (2018) investigated whether the **sex** match between lender officers and borrowers influences the likelihood that borrowers return to the same lender. They used **loan** data from a **large commercial Albanian lender** that lends money to small and medium-sized firms. Their results indicate borrowers' preferences to be attended by a person of the same sex, and that fact impacts credit market outcomes. First-time borrowers are less likely to return to lenders of the opposite sex for getting a loan. Blanco-Oliver et al. (2021) performed a similar study using data from the World Bank's Microfinance Information Exchange (MIX) platform, that records financial and operational information on micro-credit from **52 developing countries**. Their results also indicated the existence of **sex affinity** between borrowers and lender officers when deciding on a loan application. They found less discriminatory results on the overall performance of microfinance when borrowers and lenders are from the same sex (**individual fairness metric**).

Cozarenco and Szafarz (2018) investigated the existence of **sex bias** in microfinance institutions (MFI) as new barriers for female entrepreneurs to have access to micro-credit lines for their business in **France**. MFI is a government-subsidized institution looking at social performance within a budget constraint. In France, the MFI loan ceiling is EUR 10,000. Looking to broaden their clientele, but at low risk, banks are using MFI to co-finance loans for entrepreneurs without a credit history that need loans over the MFI ceiling. The authors **empirically** analyzed a dataset containing information manually fed of 1098 credit applicants from 2008 to 2012 from a **French MFI**. Their results suggest female micro-borrowers have lesser chances than male borrowers to get loans above the MFI loan ceiling. This discrimination seems to be caused by MFI transferring the clients' evaluation to the banks' assessment policy. At first look, the strategy of combining banks and MFIs to concede loans over the ceiling is discrimination neutral, but the study shows that it feeds the structural discrimination against women.

This credit discrimination against female entrepreneurs seems worldwide. In Spain, De Andrés et al. (2021) empirically evaluated a sample of **80,000 Spanish companies** looking at the **owners' sex** and their demand for credit, credit approval ratio, and credit performance. Their dataset comes from the **Spanish Central Bank (CIRBE database)**.[16] Their results showed that female entrepreneurs are less likely to have their loan application approved in their firms founding year than their male peers, comparing people from the same industry. Interestingly, they also found that women who received loans in their founding years are less

[15] Small-business loans are associated with people, not businesses.

[16] Rich in details, but confidential government data.

Table 8 Econometric studies (part1)

Ref	Bias attribute	Loan type	Country	Dataset	Finding
Beck et al. (2018)	Sex	Small-business loan	Albania	Private	Same-sex “first-time borrower"-"lender officer" influences loan approval
Blanco-Oliver et al. (2021)	Sex	Small-business loan	Developing countries	Public	Same-sex borrower-"lender officer" influences loan approval
Cozarenco and Szafarz (2018)	Sex	Small-business loan	France	Public	Female micro-borrowers have less chance than male to get loans above the microfinance institution loan ceiling
De Andrés et al. (2021)	Sex	Small-business loan	Spain	Public	Female entrepreneurs, during the founding year, are less likely to have their loan application approved than their male peers
Le and Stefańczyk (2018)	Sex	Small-business loan	Vietnam	Private	Female entrepreneurs, in male-intensive industry, are more likely to have their loans denied than their male peers
Tran et al. (2018)	Sex	Small-business loan	Vietnam	Public	Female entrepreneurs are more likely to have their loans denied than their male peers
Sackey and Amponsah (2018)	Sex	Small-business loan	Ghana	Private	No information on rejected loans undermined studies. Counterfactual analysis showed signs of structural discrimination
Sackey and Amponsah (2020)	Sex	Small-business loan	Ghana	Private	Significant presence of female entrepreneurship in Ghana, but very small participation in the credit market
Chen et al. (2017)	Sex	Any loan	China	Private	Female applicants with higher chances of getting loans in online lending platform compared to their male peers but as a higher interest rate
Maaitah (2018)	Sex	Any loan	Jordan	Public	No conclusive findings on discrimination
Li (2021)	Sex	Auto loan	India	Public	Female applicants are less likely to get a loan to buy a car than their male peers
Salgado and Aires (2018)	Sex	Any loan	Brazil	Private	Unexpected results: Female applicants have higher chances to get a loan than their male peers
Bayer et al. (2017)	Race	Mortgage	USA	Public	Black and Hispanic buyers pay more for housing regardless of the race or ethnicity of the seller
Faber (2018)	Race	Mortgage	USA	Public	Different mortgage approval rates: 71% of whites, 68% of Asians, 63% of Latinos and 54% of blacks. Additionally, black and Latino borrowers were three times more likely to receive high-cost loans compared with white applicants
Ambrose et al. (2021)	Race	Mortgage	USA	Private	Asian’ borrowers get lower interest rate than white with Asian brokers to offer lower interest rates to Asian than white borrowers. White brokers demand higher fees for Blacks, Asians, and Latinos
Bhutta and Hizmo (2021)	Race	Mortgage	USA	Public	Race and ethnicity impact interest rates, but the gaps are mitigated by discount points

likely to default. All these differences fade out as the company builds a record of profits and losses.

In Vietnam, Le and Stefańczyk (2018) **empirically** analyzed data from **national census interviews** done in 2005, 2007, 2009, 2011, 2013 with managers and employees of **small, medium, and large scale firms**. Despite the Vietnam government’s efforts to reduce sex discrimination in entrepreneurship, their research showed the likelihood of loans being denied for women-led enterprises increases to 67% in male-intensive industries and 71% in periods of tight monetary policy. Tran et al. (2018) showed similar results related to the likelihood of obtaining any loan using public data from the Vietnam Access to Resources Household Survey.

Sackey and Amponsah (2018) investigated the factors that could explain the **sex** bias against **women** in **Ghana**. They looked at the micro-credit application form randomly selected from **commercial banks micro-credit database**. There were 1,408 borrowers which received the full or a part of the requested amount, from which 872 male and 536

Table 9 Econometric studies (part2)

Refs	Bias attribute	Loan type	Country	Dataset	Finding
Loya (2022)	Race	Mortgage	USA	Public	"Black Latinos get the highest mortgage fees and mortgage denials when compared to any other white groups."
Yu (2022) & Giacoletti et al. (2021)	Race	Mortgage	USA	Public	Approval rates increase towards the end of the month. Additionally, the approval gap between **blacks and whites** drops from 7% to 3.5%
Steil et al. (2018)	Race	Mortgage	USA	Public	Higher mortgage costs for blacks and Latinos compared to whites. Collection of lawsuit case studies showing the **"reverse redlining"** effect, inducing minorities to obtain mortgages at a **high rate** (predatory mortgages) via sub-primes
Hassani (2021)	Race and Sex	Any loan	various	Academic	Race and sex discrimination are imprinted in the proxies
Park (2022)	Race and Sex	Mortgage	USA	Public	White male applicants with a white female co-applicant (e.g., a white family) have higher approval rates than any other race/sex combination
Dillbary and Edwards (2019)	Sexual orientation	Mortgage	USA	Public	Same-sex male co-applicants are significantly less likely to get their loan approved compared to heterosexual co-applicants
Sun and Gao (2019)	Sexual orientation	Mortgage	USA	Public	"Same-sex loan co-applicants are more likely to have their request denied or to get higher fees than in loans with different-sex co-applicants"
Cai et al. (2020)	No credit history	Any loan	various	Academic	Lenders increase profit while expanding the borrowers' base, by investing on getting extra information from people ranked near the denial threshold
Liu et al. (2019)	No credit history	Any loan	Low-income countries	Private	Fairness-Aware Re-ranking (FAR) algorithm balances credit assessment with borrowers fairness perspective in micro-financing
Otieno et al. (2020)	Rural jobs	Agricultural loan	various	Academic	Fuzzy, instead of a Boolean, classifiers increase the chances small-scale farmers get bank loans
Kumar et al. (2021)	Rural jobs	Agricultural loan	Developing & low-income countries	various	Small-business farmers are vulnerable groups with a low presence in bank transactions, not taking advantage of the government subsidies. ML credit scoring systems amplify bank credit eligibility reaching small-business farmers
Pi et al. (2020)	Rural jobs	Agricultural loan	China	Private	Small-business farmers get higher interest rates than city residents

Table 9 (continued)

Refs	Bias attribute	Loan type	Country	Dataset	Finding
Haupert (2022)	Technology access	Mortgage	USA	Public	Fintech lenders offer faster loan application analysis. Disparities in approval rates and pricing between Whites and non-Whites remained in Fintech lenders, but slightly lower discrimination
Fuster et al. (2019)	Technology access	Mortgage	USA	Public	Fintech lenders offer mortgage processing time 20% faster than banks, which may explain the increasing presence of Fintechs in the mortgage market that went from 2% (2010) to 8% (2016)
Allen (2019)	Technology access	Mortgage	USA	Public	Algorithmic lenders have increased risk of redlining due to learning from the biased dataset and the biased credit scoring systems

female applicants. There was no information on rejected applications. They applied counterfactual analysis, but they only found a strong correlation with **endowment amount** (**counterfactual fairness metric**). This fact could signify structural prejudice, but it would need a macro investigation. In 2020, Sackey and Amponsah (2020) extended their investigation on sex bias in the credit market domain, by looking at **loans' applications** from micro- and small-scale companies in Ghana. According to the 2019 Mastercard Index of Women Entrepreneurship,[17] **Ghana** presents the second greatest business owners percentage considering the country's all business (**demographic parity fairness metric**). In spite of this large entrepreneurship female presence, women have still timid participation in the credit market. Their research investigated the reasons to explain this paradox. Their research was based on the **empirical** analysis of questionnaire responses from 678 respondents from micro- and small-industry entrepreneurs.[18] Their results indicated that demographic characteristics including age, gender, education, size of the house and credit history, better explain the credit participation gap than sex alone.

Chen et al. (2017), using data from a Chinese online lending platform, showed a higher probability of women getting loans compared to men. However, their results showed that, although women presented a lower default rate than men, their loans were approved with higher interest rates.

Maaitah (2018) investigated factors that could explain differences in loan allocation for entrepreneur borrowers (**price differentiation**). He empirically analyzed a sample dataset containing 88,055 **Jordanian** loan applications from a **Micro Fund for Women** taking from the 2011–2017 period. He analyzed borrowers' information including sex, years of formal education, address area, and nationality. He **didn't find** any statistical explanation for the differences in the number of loans borrowed by male and female borrowers (**demographic parity fairness metric**).

Sex discrimination has not been restricted to personal, small, or micro business entrepreneurs. In India, Li (2021) developed an explanatory model using inferential statistics over a commercial financial institution dataset on auto-loan applications. Tran et al. (2018) showed similar results related to the likelihood of obtaining any type of loan using public data from the Vietnam Access to Resources Household Survey.

Contrary to the results of the main literature on discrimination, Salgado and Aires (2018) presented a study in which female entrepreneurs had better chances to get loans than their male counterparts. They used data from branches in the state of Paraíba, a northeastern state in Brazil, of a large Brazilian bank. Their results are consistent with the high female political participation in that state. They did not mention the interest rate differences.

4.5.2 Race discrimination

One of the most studied discriminatory behavior in the world is related to race or ethnicity, especially in the USA.

Bayer et al. (2017) **empirically** analyzed a very informative dataset containing a complete census of housing transactions information (**mortgage**) coming from a **private** real-estate monitoring service, (DataQuick) combined with registry information gathered under the **(HMDA)**. Data were

[17] https://www.mastercard.com/news/insights/2019/the-mastercard-index-of-women-entrepreneurs-2019/.

[18] Micro enterprise—up to 5 employees and assets not exceeding $1000.00; small enterprise—between 6 and 29 employees assets not exceeding $10000.00.

from the period between 1990 and 2008 of four **American** cities: San Francisco, Los Angeles, Chicago, and Baltimore. They were interested in price differentials in the housing market. Their results showed that black and Hispanic buyers pay more for housing regardless of the race or ethnicity of the seller. Their results suggested the estimated premia **cannot be explained by racial prejudice**, but the persistence of racial differences in home ownership, the segregation of neighborhoods, and the dynamics of wealth accumulation, i.e., signs of **structural discrimination** that are difficult to break.

Faber (2018) analyzed mortgage applications of the metropolitan statistical area housing market looking at data coming from the 2014 **(HMDA)** database. He was looking at the role of **race** in mortgages' outcomes: **mortgage approval and tax rate**. His analysis showed racial inequalities in the **USA mortgage domain**. His results showed very different **mortgage approval rates** (71% of whites, 68% of Asians, 63% of Latinos, and 54% of blacks—**demographic disparity**). Additionally, black and Latino borrowers were three times more likely to receive **high-cost loans** compared with whites (**demographic disparity**) a practice that has accelerated since the 2007–8 subprime crisis.

Ambrose et al. (2021) made an analogous study but concerning **race**. The author performed an **empirical** analysis on data of **US** mortgage loans from a large **private mortgage lender**.[19] Their data encompassed the period from January 2003 and March 2007. The dataset contained information on the borrowers' as well as the brokers' race (inferred from their names). Their result showed that Asian borrowers would get lower fees than white borrowers when dealing with Asian brokers. White brokers demand higher fees for Blacks, Asians, and Latinos. There is no evidence of price differentiation when brokers are Black or Latino.

Bhutta and Hizmo (2021) analyzed a dataset combining data from **the USA FHA*** **insured loans** originated in 2014 and 2015 with information coming from Optimal Blue Mortgage Market index[20] to identify traces of **racial and ethnic** discrimination in **mortgage pricing**. They were looking at the interest rate, discount points, and mortgage fees. They found race and ethnicity impact interest rates, but these gaps are mitigated by discount points.

Loya (2022) departs from the tri-racial stratification theory (Bonilla-Silva 2004) that considers a racial hierarchy comprising of Caucasian whites at the top, followed by honorary white[21] and the collective black group.[22] His research focused on the racial/ethnic prejudice against the Latinos community (**demographic parity**). Looking at a stratified random sample of about 115 thousand complete **mortgage** applications (single family homes) drawn from the **American HMDA** from the period of 2010 to 2017, he showed that black Latinos got the highest mortgage fees and mortgage denials when compared with other honorary white groups. This result indicates a racial prejudice even within an ethnic group by skin color.

Yu (2022) investigated the impact of algorithmic underwriting on mortgage approval behavior during the month. He looked at 25 years of **mortgage approval** data of the **American HMDA** from the period of 1994 to 2019. He accessed a **high-frequency data**, allowing a detailed monthly analysis. He observed approval rates increase towards the end of the month, reaching the highest rate on the last day of the month. He also observed that the approval gap between **blacks and whites** drops from 7% to 3.5%. A deeper analysis indicated this behavior could be explained by the incentive structure in the mortgage bank domain. Loan officers must meet their monthly loan quota, which impacts their monthly income, and non-compliance challenges their job. Further details of this approach can be found in Giacoletti et al. (2021).

Steil et al. (2018) departed from the variety of quantitative research available in the loan literature, reporting significantly higher **mortgage** costs received by blacks and Latinos compared to whites in the **USA**, to investigate the lenders' strategy to reach these borrowers (**demographic parity**). The authors performed a content analysis (**qualitative research**) of the textual depositions of **four litigation cases** against lenders whose facts were considered strong enough to hold a lawsuit and that are still ongoing cases (2022). Two cases involved the Wells Fargo Bank (the city of Baltimore vs. the Wells Fargo Bank and the City of Memphis vs. the Wells Fargo Bank), one case involved Morgan Stanley (Adkins et al. v. Morgan Stanley), and one case involved Olympia mortgage (Barkley v. Olympia Mortgage). These cases reported **"reverse redlining"** effect. Identifying minorities and using a trusting social network (such as local priests) to induce minorities to obtain mortgages at a **high rate** (predatory mortgages) via sub-primes.

[19] The "New Century Financial Corporation", which filed for bankruptcy in 2007.

[20] Optimal Blue is a private company—https://www2.optimalblue.com/obmmi/.

[21] Honorary white is defined by Bonilla-Silva (2004) as light-skinned Latinos, Japanese Americans, Korean Americans, Asian Indians, Chinese Americans, Filipinos, and most Middle Eastern Americans.

[22] Collective black is defined by Bonilla-Silva (2004) as blacks, dark-skinned Latinos, Vietnamese, Cambodians, and Laotians.

4.5.3 Both race and sex discrimination

Some studies look at the data focusing on both race and sex discrimination. Hassani (2021) used a **public dataset** from Kaggle[23] to show how algorithmic **bank loan approval** reverberates **racial/ethnic and sex** discrimination. The dataset contained 400 instances (race/ethnicity—99 "African-American", 102 "Asian", 199 "Caucasian"; sex—207 women and 193 men). 31% of the applications were rejected. Using credit scoring information, he predicted applicants' race and sex. His results suggested there will be racial and sex credit loan discrimination, even removing the sensitive attributes from the data.

Park (2022) **empirically** analyzed data from the **Federal Homing Administration** insured mortgage[24] to investigate discrimination concerning **race, ethnicity, and sex**. He analyzed over 7,6 million loans collected between 2010 and 2019, looking at rates of insurance endorsement and default among purchase mortgage applications. His results indicated that white male applicants with a female co-applicant (e.g., a white family) have higher **endorsement rates (access to service)** than any other applicant demographics with the same characteristics (**equality of opportunity fairness metric**). His findings also showed default rates did not alter the bias against Asian, Hispanic, and female applicants.

4.5.4 Sexual orientation discrimination

Loan applications do not ask for sexual orientation. Nevertheless, this information can be inferred with some statistical noise. The sex disclosure of the borrower and co-borrower can hint at that. There might be cases of father–son, mother–daughter, or friends applying for a loan. Dillbary and Edwards (2019) found evidence of sexual orientation discrimination by looking at a sample of 20% of HDMA mortgage data between 2010 and 2015. They looked at over five million mortgage applications. Their findings indicated that same-sex male co-applicants are significantly less likely to get their loan request approved compared to white heterosexual co-applicants. They also found that this type of discrimination happened across many different lenders: big or small banks and in urban or rural areas. Other studies ratify sexual orientation discrimination. Looking at the USA mortgage data, Sun and Gao (2019) showed that same-sex loan applicants and co-applicant were 73% more likely to have their request denied or to get a loan with higher fees than different-sex loan applicants.

4.5.5 No-credit-history discrimination

Sometimes, the barrier to accessing credit lines is the lack of credit history. Cai et al. (2020) addressed the discrimination against people with **no credit history**. The lack of information usually prevents good-payer applicants from receiving a loan. Gathering more information concerning the applicants' creditworthiness involves costs. Considering the lenders' goal to maximize profit and the applicants' needs for loans, the authors propose an algorithm (**in-processing prevention technique**)to identify applicants close to the decision thresholds, from which assessing more information would benefit all. They tested their ideas using a **synthetic dataset** and academic dataset (**German Credit Dataset** Hofmann 1994) containing 1,000 loan applications.

Liu et al. (2019) also addressed discrimination against poor people without a credit history. They looked at discrimination in **micro-lending** in **undeveloped countries**. They acknowledged the benefits of micro-lending for reaching people that do not have access to formal loan channels since the crowd is looking at criteria other than profit. Nevertheless, based on data from the Kiva lending platform,[25] they identified unbalanced lending among different groups of borrowers (**demographic parity fairness** metric). They propose a **Fairness-Aware Re-ranking (FAR) algorithm** to balance ranking quality and borrower-side fairness. This is an **in-processing discrimination prevention technique** to address fairness lending. They tested their algorithm for accuracy and fairness using a sample of lending transactions from 9,597 lenders taken from an 8-month period from a **proprietary Kiva dataset** .

4.5.6 Line of work discrimination

Otieno et al. (2020) discussed the discrimination against small-scale farmers to access bank loans. The author argued **small-scale farmers** cannot prove many pieces of information banks require to assess applicants' creditworthiness, such as land title and guarantees of buyers for their products. Kumar et al. (2021) did a similar study and concluded that the lack of credit history from small farmers might lead them away from mainstream banking transactions.

Pi et al. (2020) showed small farmers living in Chinese provinces get a similar discriminatory effect related to borrowers' residency location. They analyzed data form a Chinese peer-to-peer online platform (RenRen) and concluded small farmers get higher interest rates from the loans than city applicants.

[23] https://www.kaggle.com/suzanaiacob/predicting-credit-card-balance-using-regression.

[24] FHA mortgage is generally an option for borrowers with low credit score and/or small down payment amount.

[25] https://www.kiva.org/.

4.5.7 Technology-based discrimination

Considering the rapid increase of **FinTech** in the **mortgage** market, Haupert (2022) studied the impact on **racial discrimination (Individual fairness)** with the introduction of FinTechs in the mortgage market compared to traditional lenders (banks). Since the loan application and approval are online, any discrimination is more likely to be **statistical-based** and not **taste-based**. Haupert considered five racial values: Whites, Blacks, Latinos, Asians, and Others. He analyzed **HDMA data** (**American public mortgage data**) from the period of 2015 to 2017 and the neighborhood racial composition classifying into five classes according to the percentage of non-White residents (<20%, 20–40%, 40–60%,60–80%, >80%). The dataset comprised 7,630,193 first-lien mortgage applications, of which 625,474 were in the FinTech category and 6,830,311 in subprime applications, from which 529,415 were from FinTech lenders. Haupert concluded that except for Latinos, all Fintech mortgage applicants have lower mean incomes than applicants for traditional lenders. Disparities in approval rates and pricing between Whites and non-Whites remained in FinTech lenders, but slightly lower discrimination in approval rates and interest rates. When considering the neighborhood composition, Haupert concluded that predicted subprime rates increase as the neighborhood's composition of non-White residents grows, especially for Latino applicants.

Fuster et al. (2019) studied the impact of the introduction of Fintechs in the mortgage market **for borrowers with low access** to formal financial lenders, in terms of **access to loans and pricing differentiation**. The authors looked at **American** mortgage data insured by the FHA* from 2010 to 2016. Their sample dataset contained about 51,448,444 bank loans, 3,473,506 Fintech loans, and 25,604,501 loans for other lenders. The data showed a mortgage processing time 20% faster than banks and a more elastic business. These facts may explain the increasing presence of Fintechs in the mortgage market, which went from 2% (2010) to 8% (2016).

Allen (2019) **argues** that structural **race** discrimination increases by algorithmic decision-making. He highlighted the risks of algorithmic redlining caused by systems that learned from the biased dataset and the bias of current credit scoring systems as a token for creditworthiness. The high-accuracy outcomes veil social fairness performance. His argumentation focuses on **race** discrimination in the **US** housing domain, especially on **mortgage access and pricing discrimination**. He mentioned the legal actions to refrain from race discrimination in the economic environment, such as the American FHA, for broadening home ownership to a more diverse group of people, and the ECOA to fight race redlining. He claimed the need for transparency and auditing of the training datasets and the algorithms.

4.6 Human perception issues

There are two ways of looking at perception issues: human perception of the concept of fairness and human perception of the computational results' validity. Human perception of the computer outcome fairness and accuracy is fundamental for accepting and adopting ML systems. Table 10 summarizes the research findings on the human perspective of adopting automated loan approval systems. This section discusses the different approaches to dealing with human perception issues.

Many technical solutions have been proposed to lead to fair decision-making systems. However, there is no consensus on what fairness means. Wong (2020) claimed that before proposing technical solutions, it is fundamental that **the society agrees on the meaning of algorithmic fairness**. He sees this definition as a political issue. Wong grounded the discussion using **Daniels and Sabin's accountability for reasonableness framework** (Daniels and Sabin 2008), which requires the design of any algorithm to consider the interests of the affected people. Algorithm developers must consider three conditions: (a) Publicity—fairness definition, fair metrics, and the trade-offs between fairness and accuracy of any algorithmic outcome should make; (b) Acceptability—decisions should be acceptable by the affected people; (c) Revision & Appeal—conflict resolution strategies should be available.

Albach and Wright (2021) investigated the laypeople **perception of fairness across different domains**. They extend the findings of Grgić-Hlača et al. (2019) to identify which attributes were considered fair by laypeople to justify algorithmic decisions. Albach and Wright surveyed Amazon Mechanical Turk 2157 workers ("turkers"), in 2019–2020. The **questionnaire** asked turkers to rate, on a Likert scale from fair to unfair, the features used by a machine learning system to produce an outcome, as well as the agreement with the system's output. They evaluated the results using six domains, including loans, to identify moral reasoning differences. Their findings indicated that perceptions are primarily consistent across domains, except for insurance and health domains. A single dominant predictor may explain this behavior in each field that heavily impacts accuracy. Moreover, participants were turkers and not actual decision-affected people.

Binns et al. (2018) developed an empirical lab study asking 19 **UK** participants for their perception of fairness upon decisions proposed by an ML algorithm in three different scenarios, including credit loan application. Of the 19 participants, 11 were male, and eight were female. The authors were looking at five constructs: the lack of human touch in the explanation, the difficulties in interpreting the machine results, the possibility of acting upon the factors leading to the computer outcome to change it, and the relation to moral

Table 10 Human perception issues

Refs	Focus	Observations and recommendations for automated loan approval systems
Wong (2020)	People affected by automated decision-making	Recommendation for improving people acceptance of ML outcomes: (a) Publicity: make transparent the adopted fairness definition, evaluation metrics and the trade-offs between fairness and accuracy (b) Acceptability—assess affected people acceptance on computers' outcome (c) Revision & Appeal—create conflict resolution strategies
Albach and Wright (2021)	General laypeople (Mechanical Turk)	Systems' accuracy determine the human acceptance of computers' outcome. Caveat: participants were not actual automated decisions' affected people, but mechanical turk workers
Binns (2018)	General laypeople	The interface style for presenting the explanation of the machine reasoning leading to an outcome highly impacts the users' perception and acceptance of it
Saxena et al. (2019)	General laypeople (Mechanical Turk)	Acceptance of fairness definition in different application domains
Karimi et al. (2021)	People affected by automated decision-making	Recommendation for improving people acceptance of ML outcomes: provide explanations that allow people to act upon the factors preventing them to getting a loan
Rebitschek et al. (2021)	General laypeople	People underestimate ML credit scoring systems. People are more willing to accept human than computer mistakes

aspects of the machine outcomes. Their results suggest the explanation style for presenting the machine (understanding and being able to act) suggestion highly impacts the participant's perception and acceptance of the outcome.

Saxena et al. (2019) also investigated the people's perception of algorithm outcomes, but they took a different approach. They studied people's acceptance of the various definitions of fairness proposed in the algorithmic fairness literature. They build different decision-making scenarios, including loan approval and the results according to different fairness metrics (individual fairness Dwork et al. 2012, meritocracy Kearns et al. 2017; Joseph et al. 2016, and calibration Liu et al. 2017) and asked **American** Mechanical Turk participants for their perception of fairness. Their results indicated that the "calibration" definition for fairness got the best acceptance rate from the **American Mechanical Turk participants**.

Karimi et al. (2021) propose a counterfactual model that would improve people's **perception of the "feasible" actions** to change the outcome of algorithmic decision-making. For example, for a bank newcomer facing a computer outcome that jeopardized a loan application, a person may act by asking for small loans, even without needing one, to build a positive credit history and later get a bigger loan approved. An example of a non-feasible action would be "change your sex" or "get younger" to improve your credit score. The authors propose to augment counterfactual explanations to take into account the causal consequences of actions and the set of (physical) laws restricting the activities. They present a **mathematical** model for providing the minimal set of feasible measures a loan applicant can do to change the algorithm outcome.

Rebitschek et al. (2021) analyzed people's error estimations and willingness to accept errors from automated decision systems. Based on a questionnaire using 3086 respondents in Germany, they identified that people underestimate the accuracy of credit scoring systems and highly educated people are more likely to underestimate systems' errors. Surprisingly, respondents did not even accept the number of mistakes they expected the system to present in the credit scoring domain. Further, respondents were willing to take more errors from human experts than computer systems.

4.7 Definition issues

Defining decision fairness, either human or computational, is not a new issue, but it is still getting attention from researchers. Table 11 presents a summary of these studies further detailed in this section.

Researchers have been either looking to define the requirements for achieving fair algorithms (Kleinberg et al. 2016), mathematically defining metrics to measure a system's degree of fairness Kearns (2017), or even offering a programming library containing many different definitions to be able to compute and compare (Bellamy et al. 2019).

Kleinberg et al. (2016) formalized three conditions for a fair algorithm: calibration within groups (**demographic parity**), balance for the negative class across groups, and balance for the positive type across groups. They showed that the three conditions cannot be simultaneously satisfied except when you have a perfect predictor or the average prediction values for the two groups are the same.

Kearns (2017) presented various fairness metrics, highlighting the lack of consensus and conflicting definitions

Table 11 Research focus on defining fairness in ML context

References	Fairness definition	Findings
Kleinberg et al. (2016)	Requirements for fair ML systems	Conditions for fairness: (a) calibration within groups, (b) balance for the negative class across groups, and (c) balance for the positive class across groups. The three conditions cannot be simultaneously satisfied except when you have a perfect predictor
Kearns (2017)	Lack of consensus on the definition	Fairness rules should be "endogenized" into the ML training process to lead to fair outcomes
Cohen et al. (2022)	Fairness focus: on the price, on the demand, on the surplus or on the no-purchase valuation	Models for fairness in the credit domain across groups: (1) Price fairness—provide a same price (2) Demand fairness: same loan products access (3) Surplus fairness—the same difference between consumer valuation and the price paid for the loan (4) No-purchase valuation—similar average valuation of the loan across groups
Hu and Chen (2020)	Fairness focus: on price, demand, surplus or no-purchase valuation	(1) Enforcing no-purchase valuation increases social welfare. (2) Small increases in price fairness increase social welfare, but outcomes worsen for lenders and borrowers. (3) Increasing demand or surplus fairness always reduces social welfare
Lee and Floridi (2021)	Various fairness definitions in the literature	Impossibility to address all fairness definitions at the same time. Empirically showed the impact on mortgage denial for blacks when using different machine learning (ML) techniques instead of the usual logistic regression
Fuster et al. (2022)	Various fairness definitions in the literature	Nonlinear characteristic of ML techniques magnifies the differences boosting racial discrimination
Kozodoi et al. (2022)	Various fairness definitions in the literature	Cost-structure of fairness in terms of profit according to different fairness definitions and ML technique

among them. He claimed that although regulation is essential to prevent and mitigate biased algorithms, the rules should be "endogenized" into the learning process to lead to fair machine learning algorithms. His emphasis is on the need for "in-processing" solutions.

There are also efforts to define metrics for evaluating the trade-offs between results' accuracy and fairness (in the sense of diminishing discrimination against groups).

Cohen et al. (2022) present four **theoretical** models for fairness in the credit domain: fairness in price, demand, consumer surplus, and no-purchase valuation. Price fairness refers to providing a similar price (analogous to **equalized odds fairness metric**) for the goods (loans) for the groups being considered that can be classified by race, sex, or other sensitive attributes. Demand fairness refers to offering the same access to loans across groups ((analogous to **demographic parity fairness metric**)). Surplus fairness refers to having a similar difference between consumer valuation and the price paid for the loan (analogous to **equality of opportunity fairness metric**). No-purchase valuation refers to a similar average valuation of the loan across groups ((analogous to **equalized odds fairness metric**). The authors mathematically show the impossibility of being well evaluated according to the four metrics at the same time. They also present a simulation study showing the behavior of these four metrics compared to the social welfare function, as described in Hu and Chen (2020). They show that enforcing no-purchase valuation increases social welfare. They also conclude that small increases in price fairness may increase social welfare. However, as price fairness increases outcomes worsen for lenders and borrowers. Moreover, increasing demand or surplus fairness always reduces social welfare.

Lee and Floridi (2021) addresses **racial** discrimination in the **USA mortgage** domain as a trade-off analysis among possible alternative solutions for implementing algorithmic decision-making. They **empirically** showed the impact on **mortgage denial** for blacks when using different machine learning (ML) techniques, instead of the usual logistic regression. The non-linear characteristic of ML techniques magnifies the differences boosting racial discrimination, proved by Fuster et al. (2022). They also acknowledged the multitude of fairness definitions in the literature and the impossibility to address all at the same time. Based on these two facts (ML booster effect and conflicting fairness metrics), they propose to rephrase the problem and instead of having one evaluation, decision-makers should understand the trade-offs between algorithm performance, according to different ML techniques, and discrimination performance, according to a set of fairness metrics. This strategy fosters awareness of decision-makers and fosters transparency and audibility of the algorithms. They showed their method

using a dataset containing 50,000 accepted loans and 50,000 denied loans randomly taken from **2011 HMDA data**.[26] They only considered black and white borrowers. The sampled data had 90.7% white vs. black applicants. In their study they could show, for instance, a system using the Random Forest technique had the best computational performance (AUC=78%), but it would deny mortgages for almost 85% of black applicants. On the other hand, using the K-nearest neighbors classification technique would provide the second-best performance (AUC=72%), but it would deny for 65% of black applicants.

Kozodoi et al. (2022) empirically compared different machine learning techniques evaluated, considering various fairness metrics to analyze the costs in terms of profit to increase fairness. They used seven academic datasets to assess their claims.

4.8 Dataset issues

Problems in the dataset are one of the main sources of discrimination or, at least unfair computational results. Issues related to the dataset include the data source, the label trustworthiness, the missing data, the proxies, and the repair techniques. Table 12 summarizes the issues.

4.8.1 Data source issues

Technological advances create opportunities to incorporate other sources of information to assess loan applicants' credit scoring, such as cell phone payment history and social media data. According to Knight (2019), blacks are less likely to have a credit history. Including additional unstructured and semi-structured data sources improve the chances of increasing the approval rate of the **"unbanked"** minorities. Knight also advocates lighter regulation for AI systems to foster companies to innovate and to broaden ways to gather information to assess prospective clients without stereotyping. On the other hand, using informal sources of information may break people's privacy. Bryant et al. (2019) proposed using variational auto-encoder technology to create synthetic instances from the original data to preserve privacy while retaining the utility of that original data.

4.8.2 Labels

Even for people with a credit history, the dataset presents the challenge of trusting the labels assigned to the records. Chakraborty et al. (2021) claim bias and discrimination in machine learning systems come from misleading labels. They dealt with the labeling problem by proposing a pre-processing technique called Fair-SMOTE that removes records in which labels are suspected of errors. Instead of removing the records, Wakchaure and Sane (2018) proposed a **pre-processing** technique that manipulates instances considered biased according to some fairness metric. Their method comprises three steps: recognize categories and groups of examples that have been directly or indirectly discriminated, change the labels of these instances to remove bias, and use the new pre-processed dataset to train the system. They tested their approach using two **academic datasets: Adult Census Income and German credit datasets**.

4.8.3 Missing data

Not rarely, the problem is not an untrustworthy label, but a lack of labels and other attributes' values. Bogen et al. (2020) discuss the importance of collecting sensitive data, such as race and sex, to effectively combat discrimination. The authors surveyed sensitive data collection practices of **American** organizations in different domains, including credit. There is no single legal conduct. Even in the credit domain, there are different legal conducts. For instance, mortgage lenders must collect sensitive data from their borrowers and make the data public. The **HDMA dataset** contains race, ethnicity, sex, marital status, and age, among other attributes. It is not the same for consumer lenders regulated by the ECOA that prohibits gathering these data, except when used as monitoring information to avoid discriminatory behaviors.

Kallus and Zhou (2018) brought to bear that any dataset will always have some missing data and untrustworthy labels. For this reason, there will always be some degree of discrimination. They showed that even using fairness-adjusted algorithms, the **"residual discrimination"** caused by this intrinsic asymmetry of information enforces structural discrimination on the same groups focused on the fairness adjustments. They represented residual unfairness as distributions of the conditional risk score across censored and target groups. Singh et al. (2022) proposed a sampling pre-processing technique in which missing data is generated in the "neighborhood" of the minority group. Their approach is tuned to create fair classifiers for the USA mortgage domain concerning sex and race. It accounts for more than one sensitive attribute at a time. Their method was **empirically** tested using data from the HDMA national and state databases (period from 2018 and 2020).

4.8.4 Proxies

In addition to missing data and untrustworthy labels, some attributes are highly correlated with the sensitive attributes, the proxies. Even within the credit domain, there is conflicting legal guidance toward removing or maintaining sensitive

[26] The true approval rate was 75.6% in the full data set.

Table 12 Research focus on defining fairness in ML context

Refs	Dataset issues	Findings
Knight (2019)	Limited information	Blacks are less likely to have a credit history. Informal data sources may adjust the assessment of "people without a credit history's" ability and willingness to pay. On the other hand, it brings privacy concerns
Bryant et al. (2019)	Limited information	Variational auto-encoder technique to create synthetic instances to train ML systems while preserving data privacy
Chakraborty et al. (2021)	Labels' trustworthiness	Fair-SMOTE pre-processing method to identify and remove records in which labels are suspected of errors
Wakchaure and Sane (2018)	Labels' trustworthiness	Pre-processing method: (1)recognize examples indicating discrimination against a group identified by a sensitive attribute, change the labels of these examples to remove bias, and (3) use the new pre-processed dataset to train the sML system
Bogen et al. (2020)	Dataset bias	Sensitive data omission don't preclude algorithmic bias due to the existence of proxies in the dataset but hinders dataset auditing for discriminatory behavior
Kallus and Zhou (2018)	Dataset bias	There will always be a degree of discriminatory behavior. Distributions of the conditional risk score across censored and target groups represent the residual unfairness
Singh et al. (2022)	Missing data	Sampling pre-processing technique is proposed in which missing data is generated in the "neighborhood" of the minority group
Cofone (2018)	Proxies	Blocking sensitive attributes in the training data is ineffective. Measures to avoid algorithmic bias include (a) properly configuring the training dataset, (b) continuously monitoring the outcomes to detect misbehavior, and (c) creating regulations to enforce accountability
Kallus et al. (2022)	Proxies	There are many proxies leading to sex or race, such as loan applicants' surnames and addresses. Removing the protected class membership in the data may allow algorithms to infer the class membership and implicitly increase prejudice
Hort and Sarro (2021)	Proxies	Techniques to remove proxies might remove essential attributes that can lead to a distorted reality
Salimi et al. (2019)	Dataset repair	A pre-processing technique is proposed to remove instances in the dataset according to a causal pathway to outcomes that include inadmissible attributes. The causal pathway is created using extra information containing a list of admissible attributes that may impact the outcomes
Valentim et al. (2019)	Dataset repair	The performance of different pre-processing techniques in the trade-offs results between accuracy and fairness depend on domain characteristics. of pre-processing techniques

attributes. For instance, the sensitive attribute should remain when dealing with mortgages but not with micro-credit. Cofone (2018) acknowledged this **legal guidance conflict** and claimed it is **ineffective to block sensitive attributes from the training data**, given the existence of many proxies for them. He defended the benefits of modifying the sensitive characteristics in the training data using **pre-processing techniques** to avoid discrimination. Measures to avoid algorithmic bias include (a) properly configuring training set data, (b) monitoring the outcomes to detect misbehavior continuously, and (c)regulation to enforce the accountability of the person deciding with or without a decision support system.

Kallus et al. (2022) analyzed the unfairness (**equalized odds**) of algorithmic decision-making in lending money. They showed that too many proxies lead to sex or race, such as loan applicants' surnames and addresses. They looked at a sample of 14,903 **American mortgage** applications, containing only black and white applicants with annual income no more than $100,000, from the **HMDA 2011–2012 dataset**, to construct models to predict **race** from the proxy variables. They showed that inferring race from other variables might challenge even more fair lending, leading to more **mortgage denials** to minorities. They wanted to show that removing the protected class membership in the data may allow algorithms to infer the class membership and implicitly increase prejudice.

Furthermore, Hort and Sarro (2021) claimed that techniques to remove proxies that can lead to discrimination of minorities can remove essential attributes that can lead to a distorted reality. For example, students who turn in homework should perform better in exams. In this context, homework delivery should be considered an "anti-protected attribute". The authors used the academic Adult Census dataset to show that increasing the fairness of sensitive attributes prevents the discriminatory effect of anti-protected attributes. Hort and Sarro showed that grid search mitigates gender bias when using the Adult Census dataset.

4.8.5 Dataset repair

Problems in the datasets can be seen as old database problems. Salimi et al. (2019) looked at bias in datasets as a database repair problem. They focused on the bias hidden by different degrees of statistical grouping (Simpson's paradox). They proposed a **pre-processing technique** that gets, in addition to the standard list of input and output, a list of permissible attributes that may impact the outcome. The technique designates removing or including instances depending on whether it feeds a causal pathway to outcomes that include inadmissible attributes.

Valentim et al. (2019) studied the effect on fairness and performance metrics of applying different **pre-processing** techniques, such as removal of sensitive attributes, encoding of categorical features (integer encoding and one-hot encoding) and removing instances. They regarded as **statistical parity, disparate impact, and the normalized prejudice index metrics** for fairness. They used two academic datasets to perform their experiments: Adult Income and German credit data. Their results indicated that, as expected, there is no best overall pre-processing technique. Their findings suggested a high dependency on the characteristics of the domain for a trade-off analysis between fairness and outcome accuracy.

4.9 Algorithmic issues

The machine learning algorithm may introduce bias and cause morally unacceptable discrimination. The issues related to the algorithms include the amplification effect of the non-linear regression methods, classifiers' performance, knowledge representation of fairness, and the challenge of learning from imperfect labels, as summarized in Table 13

4.9.1 Amplification effect from non-linear techniques

Bono et al. (2021) **empirically** showed the outcome discrimination effect of using machine learning algorithms instead of the traditional logit credit scoring techniques. Since machine learning techniques are non-linear, slight differences are boosted. This effect not only improves accuracy performance but also increases the gap among groups. The authors investigated this effect using a **private** dataset of detailed credit data from 800,000 **UK** borrowers. Similarly, Fuster et al. (2022) showed that machine learning techniques worsened the mortgage arrangements for blacks and Latinos compared to whites using an **American** dataset of 9.37 million mortgage loans coming from 2009 to 2013 **HMDA** augmented with data from McDash, a private dataset from Black Knight company. Acknowledging the fast adoption of machine learning to evaluate creditworthiness, they mathematically proved that there is an interest rate increase as the group becomes more dispersed. They show that risky borrowers become even riskier while creditworthy borrowers become credit-worthier. Brotcke (2022) also discussed the challenges of introducing machine learning techniques to the credit marketing domain but looking at compliance issues with USA anti-discrimination laws (e.g., FHA and the ECOA).

4.9.2 Classifiers' performance issues

Otieno, Wabwoba and Musumba claim the Boolean classification between bad and good applicants leads to a significant amount of **loan rejection** for people who have difficulties proving their assets, such as the case of small farmers demonstrating their steady clientele. They proposed a **fuzzy classifier** that deals with this information imprecision. Schoeffer et al. (2021) proposed a fair ranking-based decision method that uses the relationship between legitimate features and the outcomes to compensate for the lack of such information. They tested their approach using the German credit database (**academic database**) and a synthetic dataset. They evaluated their fair ranking technique using **meritocratic unfairness** and accuracy metrics. In the examples, they looked at the **sex** sensitive attribute. They measured the cost (in terms of accuracy) to get different levels of fairness.

Instead of improving the certainty of the outcome, Coenen et al. (2020) took a different approach. They proposed a method for estimating the default in the credit domain that identifies the scenarios for which the outcome should be **"unknown"** for not being able to generate a reliable answer. Their method uses unlabeled rejected instances to improve the performance of a classifier trained with granted instances in a semi-supervised fashion. In the credit domain, they tested using two datasets: Lending Club dataset containing **public loan data** issued **between 2017 and 2018** available to borrowers and **a private dataset** of a European spot factoring company with credit lending individual invoices collected over two years.

4.9.3 Knowledge representation of fairness

Fairness and discrimination have been mainly represented as patterns inferred by data. Cai et al. (2020) represented fairness as a criterion that lenders should consider in this resource allocation problem, besides profit maximization. They proposed an algorithm to identify applicants on the border of having their loan application approved, close to the decision threshold. Lenders should consider spending some cash gathering extra information about these borderline applicants. The authors show it is worthwhile in terms of the return on lenders' investment. Elzayn et al. (2019) also rephrase the discrimination problem of **algorithmic decision fairness** as a problem of resource allocation, including

Table 13 Summary of the algorithmic Issues

Refs	Issue	Findings
Bono et al. (2021)	Effects of non-linear techniques on loan approval outcomes	ML emphasizes discrimination due to slight differences in data boosts differences in the outcomes. This effect improves accuracy performance but also increases the gap among groups
Fuster et al. (2022)	Effects of non-linear techniques on loan approval outcomes	Mathematical proof that the interest rate increases as the group becomes more dispersed. They show that risky borrowers become even riskier while creditworthy borrowers become credit-worthier
Brotcke (2022)	Effects of using machine learning techniques on loan approval outcomes	Challenges to make ML systems comply to USA anti-discrimination laws
Otieno et al. (2020)	Effects of Boolean classifiers in loan approval	Fuzzy systems improve in terms of fairness loan application analysis
Schoeffer and Kuehl (2021)	Imperfect labels	Fair ranking-based decision method that uses the relationship between legitimate features and the outcomes to compensate for unfair results
Coenen et al. (2020)	Loan default estimation performance	Increase ML reliability by a loan default predictor that identifies the scenarios for which the outcome should be **"unknown"**
Cai et al. (2020)	Knowledge representation	Fairness should be included an additional criterion in loan approval algorithm
Elzayn et al. (2019)	Knowledge representation	Algorithmic fairness as a resource allocation problem including fairness as an additional optimization criterion
Farnadi et al. (2018)	Knowledge representation	symbolic representation of fairness comes from the logic domain
Kallus and Zhou (2018)	Imperfect labels	Even using fairness-adjusted algorithms, the "residual discrimination" caused by intrinsic information asymmetry enforces structural discrimination
Ghosh et al. (2021)	Imperfect labels	The use of inferred labels from demographic information can increase unfair results

equality of opportunity as an additional criterion to be considered. They model the problem and allow measuring the cost for fairness by measuring the solution's utility considering the available resources. The resource allocation algorithm starts with an unknown distribution of the candidates in each group needing resources. At each round, the algorithm allocates resources, so individuals from any group have similar probabilities of receiving resources. The allocation is evaluated, and feedback is considered by the learning algorithm, adjusting the allocation behavior for the next round. This **in-processing** approach has received a great deal of attention because it circumscribes the algorithmic fairness problem into a well-known area of resource allocation with multiple objective criteria.

Another approach toward a symbolic representation of fairness comes from the logic domain. Farnadi et al. (2018) proposed a machine learning algorithm for relational datasets that take into account fairness patterns as first-order logic axioms (**in-processing algorithmic technique**). They tested their framework (FairPSL) using **synthetic data**, evaluating result accuracy and fairness. The authors suggested their approach can lead to both accurate and fair decisions.

4.9.4 Learning from imperfect labels

Lack of data labels and uncertainties are unavoidable problems that the algorithm must deal with. Kallus and Zhou (2018) brought to bear the **discrimination** issue caused by **improper, but the only feasible, data collection**. For instance, in the credit loan domain, loan default is only observed on approved loan applicants and used to train machine learning credit loan systems, perpetuating discrimination. The authors showed that even using fairness-adjusted algorithms, the **"residual discrimination"** caused by this intrinsic information asymmetry enforces structural discrimination on the same groups focused on the fairness adjustments. They represented residual unfairness as distributions of the conditional risk score across censored and target groups.

Frequently, pre-processing techniques are insufficient to avoid this challenge to the machine learning algorithm. Moreover, some methods for handling missing labels may worsen the problem. For example, (Ghosh et al. 2021) have shown that using inferred labels from demographic information can even increase unfair results. When automatically

obtaining values for sensitive attributes from people's photos, names, and addresses, the inference errors lead to mistakes that must be accounted for but rarely are.

4.10 Outcome issues

Last but not least, as summarized in table 14, there are issues related to the outcome of intelligent systems that varies from mistaken outcomes to practical explanation to allow users to overcome barriers in future applications and governance of algorithmic decision-making based on the results.

Lohia et al. (2019) proposed a method to prioritize specific instances (instances' weights) to change the classifiers' outputs. Similarly to Kamiran and Calders (2012) that select instances to change the classifier's outcome, Lohia et al. propose to change the instances more likely to be biased considering sex, race, and age. They tested their approach using **academic datasets** including the **German credit dataset**. They analyzed **sex, race, and age** bias separately.

Karimi et al. (2021) claimed that instead of acting solely on the computational side, humans should be more active by understanding the system's outcomes. In the case of a loan, an excellent explanation for rejection should include elements for which people could act upon changing their chances of getting a positive result in a future loan application.

Mendes and Mattiuzzo (2022) discussed the governance of algorithmic decision-making in the **credit scoring** domain in the light of current **Brazilian legislation**. They focused on discrimination caused by statistical error, generalization, use of sensitive information, and inadequate correlation. The literature indicates transparency and accountability as essential strategies to combat algorithmic discrimination. Nevertheless, because of business confidentiality issues, the authors do not believe transparency is feasible in the credit scoring domain. On the other hand, accountability can be enforced by legislation. The Brazilian general data protection act helps to move forward, but it is open to different interpretations, depending on the consistent and firm action of the Data Protection Authority.

5 Conclusion and open issues

Although fair algorithms and discrimination-free decision-making have been intensely depicted by current research in the data-driven financial domain, there are still many unexplored areas that deserve attention. The section addresses the most substantial ones among them, summarized in Fig. 10.

5.1 Broadening discrimination scope

Most papers addressed prejudice, in the loan applications, against blacks (racial discrimination), Latinos (ethnicity discrimination), women (sex discrimination), or general "unfair results". There are few papers that addressed prejudice concerning their line of work, such as small farmers (Otieno et al. 2020; Pi et al. 2020), people without a credit history (Cai et al. 2020; Liu et al. 2019) and sexual orientation (Sun and Gao 2019; Dillbary and Edwards 2019). Moreover, research on prejudice against certain sexual orientations is restricted to mortgages due to data gathering difficulty. At the same time that gathering more information on sensitive attributes such as sexual orientation and physical disabilities may feed algorithmic prejudices, this information can also help to identify and monitor prejudices in human or computational decision-making. Another important observation from the papers is the need for regulation.

5.2 Analyze multiple sensitive attributes together

Most studies have studied the discrimination effects by looking at one single sensitive attribute at a time. Still, as explained by Crenshaw's intersectionality theory (Crenshaw 1989), sex and race united bring a stronger form of discrimination. Dillbary (Dillbary and Edwards 2019) considered the combination of race and sexual orientation, discrimination against black male homosexuals leads to loan rejection higher than white heterosexual males. On the algorithmic side, Singh et al. (2022) proposed a

Table 14 The issues related to the outcome of machine learning systems

Refs	Issue	Findings
Lohia et al. (2019)	Bias on outcomes	Post-processing method that looks at a bias to change the weight of specific instances in the ML training phase to adjust the outcomes
Kamiran and Calders (2012)	Algorithmic outcome explanation	Human role on auditing ML systems calls for actionable explanation to enable different outcomes
Mendes and Mattiuzzo (2022)	Governance of algorithmic decision-making	They focused on discrimination caused by statistical errors. Algorithmic reasoning transparency and decision-making accountability are essential strategies to combat algorithmic discrimination

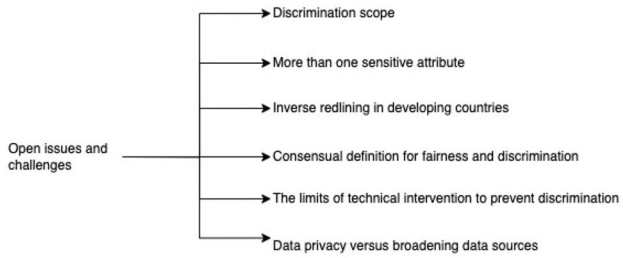


Fig. 10 Research opportunities

method to consider more than one sensitive attribute at a time. Since discrimination is usually broader than towards a single factor, there is a call for more research considering multiple attributes at a time.

5.3 Reverse redlining challenge

For a long time, there have been studies showing indirect racial discrimination considering the person's home address (Black 1999). This fact could be crucial in banks' decisions on loan applications. A related distortion is that banks and other financial institutions started using this information to push loan offers at a high rate disguised as a good opportunity. This reverse redlining has been studied in the USA mortgage domain (Steil et al. 2018).

Redlining and reverse redlining constitute a paradox. People are denied a mortgage due to some specific attribute value (e.g., race) but are also targeted through marketing campaigns to get a mortgage in much worse conditions regarding pricing, i.e., a higher interest rate. Swan (2019).

Understanding redlining and reverse redlining effects concerning race is still more complicated in the global south countries. Taking into account Bourdieu's theory of skin color as a symbolic asset in society (Prasad 2022; Bourdieu 2018) argues that, while in countries with a white majority, race is the factor of prejudice, in the global south countries, in which there is a strong miscegenation, skin tone, and facial attributes that resemble European characteristics are the determining factors. For this reason, it is more difficult to pinpoint the problem. People with lighter skin tones within the racial categorization have better chances to get a loan.

There is a need for studies on reverse redlining in the global south countries (developing countries) using large datasets that could enlighten the effect of the country's development and inequality on credit provision.

5.4 Consensus on fairness definition

There are many definitions for algorithmic fairness and discrimination, as presented in previous sections. There are even computational libraries to calculate the fairness degree according to many metrics (Bellamy et al. 2019). Nevertheless, Segal et al. (Segal et al. 2021) proposed a fairness certification for machine learning systems, focusing on the training dataset. For maintaining data confidentiality, they proposed to use cryptography features to be able to check the machine learning training dataset.

Economists and operational researchers would know-how to model the context once they understand the scope and context. Computer scientists have demonstrated they know-how to implement fair algorithms and prevent discrimination, as long as they have a clear definition of the problem: what is fair? what is unacceptable discrimination? Analogously, regulators know-how to write laws to restrict undesired behaviors as the problem becomes clear.

On the other hand, a consensus definition that is accepted by society is needed. Moreover, since culture influences society's perception, probably the definition is not general worldwide. Looking at the differences in accepted definitions and the impact on decisions and systems is a challenge.

5.5 Technical limitations

A fair algorithm, for any definition, does not guarantee discrimination-free results. Society carries many inequalities that are very persistent. Sometimes, discrimination is reinforced by the laws, for example, in the case of the old USA mortgage law.

Technically, we can offer the same opportunities to people under the very same conditions. However, the number of people in that specific conditions might be very different across groups. Consequently, it is not a matter of being fair, but it involves costs for society that must be discussed.

Technical action has boundaries. Going over these boundaries to deal with structural discrimination may call for government affirmative actions and/or laws. The American FHA and the ECOA are examples of government actions to address structural discrimination. Laws have a positive effect on combating discrimination. Dillbary showed in his study of mortgage data from 2005 to 2015 (Dillbary and Edwards 2019), discrimination against homosexual mortgage applicants was much lower in the USA states that had passed anti-sexual orientation discrimination laws. In 2017, the USA passed a federal law prohibiting sexual discrimination in mortgage deliberation that changed the scenario.

5.6 Data privacy versus data sources' widening

New loan applicants have been discriminated against for the lack of information that would help lenders to assess the risk of approving a loan application. Even when the credit history is not available, applicants may have cell phone payment

records. Knight (2019) did a preliminary study using this type of information. But where does an individual's privacy stand? Who does it help: lenders or applicants? Should it be indiscriminately used? What are the consequences? Have these data been used in a real scenario? There are international laws, such as the GDPR[27] and CCPA,[28] protecting the use of personal data without consent. What are the ethical concerns here? These are some of the questions that must be addressed before widening the data sources to evaluate loan applicants.

5.7 Final remarks

This paper presents comprehensive literature focusing on studies on discrimination in the credit domain using a systematic review method considering five data sources (ACM Digital Library, IEEE Digital Library, Scopus, Springer Link Google Scholar). This review was conducted by three researchers that examined and categorized the existing literature. Out of the 1320 initial research papers located in the data sources, 78 papers were selected.

The main threats affecting the validity of our SLR are related to the way we selected, extracted, and filtered the papers used for our analyses. We mitigate these threats by defining a search strategy that uses expressions and synonyms and alternative spelling and by using two well-known tools, PARSIF and Publish or Perish, to avoid human unconscious skipping.

Our analyses bring evidence of US dominance in the research concerning discrimination in the credit domain. It also shows that most of the current research is conducted via econometric analysis of existing mortgage datasets, mostly the public HDMA. Results have shown the existence of discrimination in many countries, against women, blacks, Latinos, and male homosexuals. Both loan rejection and price differentiation were found. Algorithmic discrimination has been taking a more general view to avoid unfair results. Moreover, there is no consensual definition of algorithmic fairness, and the existent metrics can lead to contradictions. Our analyses also reveal open issues and opportunities for future research.

Based on our findings, we argue there is still a wide room for further research improvement. Moreover, the fast widespread of machine learning in the credit domain allied to more strict discrimination laws makes further research even more necessary. The open issues that emerged in this study may represent the input for researchers interested in developing more powerful techniques for combating perceived lack of fairness, identifying discrimination, preventing algorithmic discrimination, and acting upon decisions' outcomes to change the scenario in a future loan application.

As AI becomes more and more pervasive in the provision of many services, it is of utmost importance to have a code of conduct that assures both the users and the service providers that discrimination, in its several guises (race, gender, age, etc.), is being avoided as much as possible.

The principles and procedures reviewed in this paper are applicable to other areas. Discrimination exists in the provision of health, education, security, supply of utilities (such as water, electricity, and phone services), and general discrimination in the provision of consumption good services. In all of these areas, discrimination takes place in one of the following three dimensions: the intensive margin, the extensive margin, and the quality margin. Not all forms of discrimination exist in all of them, the credit market is perhaps one of the few where all aspects exist.

The intensive margin refers to how discrimination affects the amount of service provided. In the case of credit, it is about the size of the loan, in the case of security it is about the amount of policing that takes place, and in health care is about the time devoted by the doctor attending certain types of patients. The extensive margin is not about the marginal change but the exclusion from access. So, in the credit market, it is about whether or not a person has access to credit, whether or not an individual has access to education or health care, to whether or not a certain set of individuals have more disruption of services such as water, electricity, and phone, or whether or not a store determines whether or not to accept a person into the store depending on who the person is or looks like. Finally, the quality margin refers to the fact that the quality of the service is related to individual characteristics in an unreasonable way. For example, certain individuals receive lower quality of health care and education, and in the credit market, credit management is worse (or the loan conditions are worse).

The advantage of studying the credit market is that all forms of discrimination are prevalent, and all three are extremely costly to citizens: i.e., not getting the quantity needed, not getting the loan at all, or once the loan is obtained that treatment is worse. By concentrating on the credit market, we can extrapolate and understand discrimination in other economic activities.

Appendix A

Tables 15, 16, 17, 18 present the research settings of the 78 analyzed papers.

[27] General Data Protection Regulation—https://gdpr-info.eu/.

[28] California Consumer Privacy Act of 2020—https://oag.ca.gov/privacy/ccpa.

Table 15 Research settings (part 1). Abbreviations: "first author": M = male and F = female, A = Asian, B = black, I = Indian, L = Latino, W = white; "Impact": RL = rejected loan application, DF = differentiation on pricing, UR = unfair results; "Publication domain": CS = computer science, PHY = philosophy or social science, ECO = economics/business, Law = law

Id	Refs	Area	First author info	Research type	Research topic	Discrimination grounding theory	Dataset type	Country	Impact
1	(Albach and Wright 2021)	CS	F-W	Empirical	Fairness Perception	Implicit discrimination	Own dataset	USA	RL
2	(Bellamy et al. 2019)	CS	F-W	Theoretical	Formulation	Implicit discrimination		USA	UR
3	(Binns et al. 2018)	CS	M-W	Empirical	Fairness Perception	Implicit discrimination	Own dataset	UK	UR
4	(Bogen et al. 2020)	CS	F-W	Essay	Data gathering	Implicit discrimination		USA	UR
5	(Bryant et al. 2019)	CS	M-B	Theoretical	Privacy	Implicit discrimination		USA + PT	UR
6	(Cai et al. 2020)	CS	M-A	Explanatory	No-credit history	Implicit discrimination	Academic		RL
7	(Chakraborty et al. 2021)	CS	M-I	Empirical	Dataset bias	Implicit discrimination	Academic		UR
8	(Chen et al. 2017)	CS	M-A	Explanatory	Sex bias	Implicit discrimination	Private	China	RL
9	(Coenen et al. 2020)	CS	F-W	Empirical	Default estimation	Implicit discrimination	Public + Private	USA + EU	RL
10	(Cohen et al. 2022)	CS	M-W	Theoretical	Fairness definition	Implicit discrimination		USA	UR
11	(Corrales-Barquero et al. 2021)	CS	M-L	Essay	Literature review				UR
12	(Elzayn et al. 2019)	CS	M-W	Theoretical	Algorithmic fairness	Implicit discrimination			UR
13	(Farnadi et al. 2018)	CS	F-I	Empirical	Fairness representation	Implicit discrimination	Synthetic		UR
14	(Ghosh et al. 2021)	CS	M-I	Empirical	Inference uncertainties	Implicit discrimination	Academic		UR
15	(Hassani 2021)	CS	M-I	Empirical	Race & sex estimation from data	Implicit discrimination	Academic		UR
16	(Hort and Sarro 2021)	CS	M-W	Empirical	Proxies	Information asymmetry	Academic		UR
17	(Kallus and Zhou 2018)	CS	M-W	Essay	Residual discrimination	Information asymmetry	Academic		UR
18	(Kallus et al. 2022)	CS	M-W	Explanatory	Proxies	Information asymmetry	Public (HDMA)	USA	RL
19	(Karimi et al. 2021)	CS	M-I	Theoretical	Useful explanation	Required skills Implicit discrimination			UR
20	(Kearns 2017)	CS	M-W	Theoretical	Metrics	Implicit discrimination			UR

Table 16 Research Settings (Part2). Abbreviations: "first author": M = male and F = female, A = Asian, B = black, I = Indian, L = Latino, W = white; "Impact": RL = rejected loan application, DF = differentiation on pricing, UR = unfair results; "Publication domain": CS = computer science, PHY = philosophy or social science, ECO = economics/business, Law = law

Id	Refs	Area	First author info	Research type	Research topic	Discrimination grounding theory	Dataset type	Country	Impact
21	(Kleinberg et al. 2016)	CS	M-W	Theoretical	Metrics	Implicit discrimination			UR
22	(Kordzadeh and Ghasemaghaei 2022)	CS	M-W	Essay	Impact on decision	stimulus-organism-response			UR
23	(Lee and Floridi 2021)	CS	F-A	Empirical	Fairness trade-offs	Statistical Model	Public (HDMA)	USA	RL
24	(Liu et al. 2019)	CS	M-A	Empirical	Online micro-credit	Implicit discrimination	Private (Kiva)	Low-income countries	RL
25	(Lohia et al. 2019)	CS	M-I	Empirical	Outcome fairness	Implicit discrimination	Academic		UR
26	(Mehrabi et al. 2021)	CS	F-I	Essay	Literature review	Implicit discrimination			UR
27	(Moscato et al. 2021)	CS	M-W	Empirical	Unbalanced datasets	Implicit discrimination	Public	USA	RL
28	(Ragnedda 2020)	CS	M-W	Essay	Inequalities definition				UR
29	(Salimi et al. 2019)	CS	M-I	Empirical	Database repair	Statistical model	Academic		UR
30	(Saxena et al. 2019)	CS	F-I	Empirical	Fairness perception	Implicit discrimination	Own dataset	USA	UR
31	(Schoeffer et al. 2021)	CS	M-W	Empirical	Fairness trade-offs	Implicit discrimination	Academic		UR
32	(Segal et al. 2021)	CS	M-W	Empirical	Fairness certification	Implicit discrimination	Academic		UR
33	(Singh et al. 2022)	CS	M-I	Empirical	Missing data	Information asymmetry	Public (HDMA)	USA	UR
34	(Sun and Gao 2019)	CS	M-A	Explanatory	Sexual orientation bias	Implicit discrimination	Public (HDMA)	USA	RL + DP
35	(Valentim et al. 2019)	CS	F-W	Empirical	Pre-processing comparison	Taste-based	Academic		UR
36	(Wakchaure and Sane 2018)	CS	M-I	Empirical	Bias on instances	Information asymmetry	Academic		RL
37	(Ambrose et al. 2021)	ECO	M-W	Explanatory	Race/ethnicity bias	Taste-based	Private	USA	DP
38	(Aitken 2017)	ECO	M-W	Essay	Literature review	Visual legibility			
39	(Bayer et al. 2017)	ECO	M-W	Explanatory	Structural discrimination	Taste-based	Public & Private	USA	DP
40	(Beck et al. 2018)	ECO	M-W	Explanatory	Sex bias	Taste-based	Private	Albania	DP

Table 17 Research Settings (Part 3). Abbreviations: "first author": M = male and F = female, A = Asian, B = black, I = Indian, L = Latino, W = white; "Impact": RL = rejected loan application, DF = differentiation on pricing, UR = unfair results; "Publication domain": CS = computer science, PHY = philosophy or social science, ECO = economics/business, Law = law; Dev* = Developing; Race = Race or ethnicity

Id	Refs	Area	First author info	Research type	Research topic	Discrimination grounding theory	Dataset type	Country	Impact
41	(Bhutta and Hizmo 2021)	ECO	M-I	Explanatory	Race bias	Taste-based	Public & Private	USA	DP
42	(Black 1999)	ECO	M-B	Essay	Redlining	Taste-based		USA	RL + DP
43	(Blanco-Oliver et al. 2021)	ECO	M-W	Explanatory	Sex bias	Taste-based	Public	Dev* countries	RL
44	(Bono et al. 2021)	ECO	F-W	Explanatory	Sex & Race	Statistical model	Private	UK	RL
45	(Brotcke 2022)	ECO	F-W	Essay	Law's compliance	Statistical model		USA	RL
46	(Cozarenco and Szafarz 2018)	ECO	F-W	Explanatory	Sex bias	France	Taste-based	France	RL
47	(De Andrés et al. 2021)	ECO	M-W	Explanatory	Sex bias	Taste-based	Private	Spain	RL
48	(Faber 2018)	ECO	M-B	Explanatory	Race bias	Taste-based	Public (HDMA)	USA	RL + DP
49	(Fuster et al. 2019)	ECO	M-W	Explanatory	Fintech	Taste-based	Public (HDMA)	USA	RL
50	(Fuster et al. 2022)	ECO	M-W	Explanatory	Unfair results	Statistical model	Public (HDMA)	USA	DP
51	(Giacoletti et al. 2021)	ECO	M-W	Explanatory	Race bias	Taste-based	Private	USA	RL
52	(Le and Stefańczyk 2018)	ECO	M-A	Explanatory	Sex bias	Taste-based	Public	Vietnam	RL
53	(Li 2021)	ECO	F-A	Explanatory	Sex bias	Taste-based	Private (auto-loan)	India	RL + DP
54	(Loya 2022)	ECO	M-L	Explanatory	Race bias	Tri-racial stratification	Public (HDMA)	USA	RL
55	(Maaitah 2018)	ECO	M-I	Explanatory	Sex bias	Taste-based	Microcredit	Jordan	RL
56	(Mitchell et al. 2021)	ECO	F-W	Essay	Fairness definition		Required skills		UR
57	(Nyarko 2022)	ECO	M-B	Explanatory	Sex bias	Taste-based	Public (MFI)	Dev* countries	RL + DP
58	(Otieno et al. 2020)	ECO	M-B	Explanatory	Line of work bias	Required skills	Academic		RL
59	(Park 2022)	ECO	M-W	Explanatory	Sex & Race	Final outcome	Public (FHA)	USA	RL
60	(Pi et al. 2020)	ECO	M-A	Explanatory	Line of work bias	Required skills	Private (Ren-Ren)	China	RL + DP

Table 18 Research Settings (Part 4). Abbreviations: "first author": M = male and F = female, A = Asian, B = black, I = Indian, L = Latino, W = white; "Impact": RL = rejected loan application, DF = differentiation on pricing, UR = unfair results; "Publication domain": CS = computer science, PHY = philosophy or social science, ECO = economics/business, Law = law

Id	Refs	Area	First author info	Research type	Research topic	Discrimination grounding theory	Dataset type	Country	Impact
61	(Rebitschek et al. 2021)	ECO	M-W	Empirical	Error perception	Visual legibility	Own dataset	Germany	UR
62	(Sackey and Amponsah 2018)	ECO	M-B	Explanatory	Sex bias	Final outcome	Private	Ghana	RL
63	(Sackey and Amponsah 2020)	ECO	M-B	Explanatory	Sex bias	Final outcome	Own dataset	Ghana	RL
64	(Salgado and Aires 2018)	ECO	F-L	Explanatory	Sex bias	Taste-based	Private	Brazil	DP
65	(Steil et al. 2018)	ECO	M-W	Explanatory	Reverse redlining	Taste-based	Own dataset	USA	DP
66	(Tran et al. 2018)	ECO	M-A	Explanatory	Sex bias	Taste-based	Public	Vietnam	RL
67	(Yu 2022)	ECO	M-A	Explanatory	Race bias	Taste-based	Public (HDMA)	USA	RL
68	(Allen 2019)	Law	M-W	Race bias	Race bias	ECOA, FHA	Public & Private	USA	RL
69	(Bruckner 2018)	Law	M-W	Essay	Lender 2.0	USA laws		USA	RL
70	(Cofone 2018)	Law	M-L	Essay	Proxies	USA laws		USA	UR
71	(Dillbary and Edwards 2019)	Law	M-W	Explanatory	Sexual orientation bias	Intersectionality	Public (HDMA)	USA	RL + DP
72	(Knight 2019)	Law	M-A	Essay	Data gathering paradox	GDPR, CCPA			RL
73	(Mendes and Mattiuzzo 2022)	Law	F-L	Essay	Governance	LGPD		Brazil	UR
74	(Swan 2019)	Law	F-W	Essay	Reverse redlining	USA laws		USA	RL + DP
75	(Baesens et al. 2003)	OR	M-W	Empirical	Credit scoring comparison	FICO	Academic		UR
76	(Kozodoi et al. 2022)	OR	M-W	Empirical	Credit scoring comparison	FICO	Academic		UR
77	(Prasad 2022)	PHY	M-A	Essay	Skin tone bias	Symbolic capital	Taste-based	Global south	RL
78	(Wong 2020)	PHY	M-A	Essay	Literature review	Taste-based			UR

Author Contributions All authors contributed to the study's conception and design. Material preparation, data collection, and the first draft of the manuscript were performed by Ana C.B. Garcia. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript. Marcio G.P. Garcia and Roberto Rigobon reviewed and supervised the paper writing.

Funding Open Access funding provided by the MIT Libraries. The authors have no relevant financial or non-financial interests to disclose.

Data availability Not applicable.

Declarations

Conflict of interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Act CR (1978) PART 1607: Uniform Guidelines on Employee Selection Procedures (1978). https://www.govinfo.gov/content/pkg/

CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml
Aitken R (2017) All data is credit data: Constituting the unbanked. Competition & Change 21(4):274–300
Albach M, Wright JR (2021) The role of accuracy in algorithmic process fairness across multiple domains. In: Proceedings of the 22nd ACM Conference on Economics and Computation, pp 29–49
Alesina AF, Lotti F, Mistrulli PE (2013) Do women pay more for credit? Evidence from Italy. J Eur Econ Assoc 11(suppl–1):45–66
Allen JA (2019) The color of algorithms: An analysis and proposed research agenda for deterring algorithmic redlining. Fordham Urb. LJ 46:219
Alliance NFH (2014) Zip Code Inequality: Discrimination by Banks in the Maintenance of Homes in Neighborhoods of Color. https://nationalfairhousing.org/wp-content/uploads/2022/02/2014_08_27_NFHA_REO_report.pdf
Ambrose BW, Conklin JN, Lopez LA (2021) Does borrower and broker race affect the cost of mortgage credit? The Review of Financial Studies 34(2):790–826
Arrow KJ (2015) The theory of discrimination. In: Discrimination in Labor Markets, pp 1–33. Princeton University Press, Princeton
Atkins R, Cook L, Seamans R (2022) Discrimination in lending? Evidence from the paycheck protection program. Small Bus Econ 58(2):843–865
Aztiria A, Izaguirre A, Basagoiti R, Augusto JC, Cook DJ (2010) Automatic modeling of frequent user behaviours in intelligent environments. In: 2010 Sixth International Conference on Intelligent Environments, pp 7–12. IEEE
Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J, Vanthienen J (2003) Benchmarking state-of-the-art classification algorithms for credit scoring. Journal of the operational research society 54(6):627–635
Banasik J, Crook J (2007) Reject inference, augmentation, and sample selection. Eur J Oper Res 183(3):1582–1594
Barocas S, Selbst AD (2016) Big data's disparate impact. Calif L Rev 104:671
Bartlett R, Morse A, Stanton R, Wallace N (2022) Consumer-lending discrimination in the fintech era. J Financ Econ 143(1):30–56
Bayer P, Casey M, Ferreira F, McMillan R (2017) Racial and ethnic price differentials in the housing market. J Urban Econ 102:91–105
Beck T, Behr P, Madestam A (2018) Sex and credit: Is there a gender bias in lending? Journal of Banking and Finance 87
Becker GS (2010) The Economics of Discrimination. University of Chicago Press, Chicago
Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilović A et al (2019) Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Dev 63(4/5):4–1
Berk R, Heidari H, Jabbari S, Kearns M, Roth A (2021) Fairness in criminal justice risk assessments: the state of the art. Sociol Methods Res 50(1):3–44
Bertrand M, Chugh D, Mullainathan S (2005) Implicit discrimination. American Economic Review 95(2):94–98
Bhutta N, Hizmo A (2021) Do minorities pay more for mortgages? The Review of Financial Studies 34(2):763–789
Binns R (2018) Fairness in machine learning: Lessons from political philosophy. In: Conference on Fairness, Accountability and Transparency, pp 149–159. PMLR
Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems, pp 1–14
Black HA (1999) Is there discrimination in mortgage lending? what does the research tell us? Rev Black Polit Econ 27(1):23–30
Blanco-Oliver A, Reguera-Alvarado N, Veronesi G (2021) Credit risk in the microfinance industry: The role of gender affinity. J Small Bus Manage 59(2):280–311
Bogen M, Rieke A, Ahmed S (2020) Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 492–500
Bonilla-Silva E (2004) From bi-racial to tri-racial: Towards a new system of racial stratification in the usa. Ethn Racial Stud 27(6):931–950
Bono T, Croxson K, Giles A (2021) Algorithmic fairness in credit scoring. Oxf Rev Econ Policy 37(3):585–617
Bourdieu P (2018) Distinction: A social critique of the judgment of taste. In: Grunsky DB (ed) Social Stratification. Routledge, London, pp 982–1003
Brotcke L (2022) Time to assess bias in machine learning models for credit decisions. Journal of Risk and Financial Management 15(4):165
Bruckner MA (2018) The promise and perils of algorithmic lenders' use of big data. Chi.-Kent L. Rev. 93:3
Bryant R, Cintas C, Wambugu I, Kinai A, Diriye A, Weldemariam K (2019) Evaluation of bias in sensitive personal information used to train financial models. In: 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp 1–5. IEEE
Cai W, Gaebler J, Garg N, Goel S (2020) Fair allocation through selective information acquisition. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 22–28
Calmon F, Wei D, Vinzamuri B, Natesan Ramamurthy K, Varshney KR (2017) Optimized pre-processing for discrimination prevention. Advances in neural information processing systems 30
Chakraborty J, Majumder S, Menzies T (2021) Bias in machine learning software: why? how? what to do? In: Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, pp 429–440
Charles KK, Hurst E, Stephens M (2008) Rates for vehicle loans: race and loan source. Am Econ Rev 98(2):315–20
Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. J Artif Intell Res 16:321–357
Chen D, Li X, Lai F (2017) Gender discrimination in online peer-to-peer credit lending: evidence from a lending platform in china. Electron Commer Res 17(4):553–583
Chen I, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? Advances in Neural Information Processing Systems 31
Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163
Coenen L, Abdullah AK, Guns T (2020) Probability of default estimation, with a reject option. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), pp 439–448. IEEE
Cofone IN (2018) Algorithmic discrimination is an information problem. Hastings LJ 70:1389
Cohen MC, Elmachtoub AN, Lei X (2022) Price discrimination with fairness constraints. Management Science
Colquitt JA, Rodell JB (2015) Measuring justice and fairness. In: Cropanzano R, Ambrose ML (eds) The Oxford Handbook of Justice in the Workplace. Oxford University Press, Oxford, pp 187–202 (**Chap. 8**)
Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv preprint arXiv:1808.00023
Corrales-Barquero R, Marín-Raventós G, Barrantes EG (2021) A review of gender bias mitigation in credit scoring models. 2021

Ethics and Explainability for Responsible Data Science (EE-RDS), 1–10
Cozarenco A, Szafarz A (2018) Gender biases in bank lending: Lessons from microcredit in france. J Bus Ethics 147(3):631–650
Crenshaw K (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics. the university of chicago legal forum, 1989 (1), 139-167. Chicago, IL
Crenshaw K (1991) Mapping the margins: Intersectionality, identity politics, and violence against women of color. Stanford Law Review 43(6):1241–1299
Daniels N, Sabin JE (2008) Accountability for reasonableness: an update. BMJ 337
Datta A, Fredrikson M, Ko G, Mardziel P, Sen S (2017) Proxy non-discrimination in data-driven systems. arXiv preprint arXiv:1707.08120
De Andrés P, Gimeno R, Cabo RM (2021) The gender gap in bank credit access. J Corp Finan 71:101782
Dikmen M, Burns CM (2016) Autonomous driving in the real world: Experiences with tesla autopilot and summon. In: Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp 225–228
Dillbary JS, Edwards G (2019) An empirical analysis of sexual orientation discrimination. The University of Chicago Law Review 86(1):1–76
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
Elegido JM (2011) The ethics of price discrimination. Bus Ethics Q 21(4):633–660
Elhassan T, Aljurf M (2016) Classification of imbalance data using tomek link (t-link) combined with random under-sampling (rus) as a data reduction method. Global J Technol Optim S 1
Elzayn H, Jabbari S, Jung C, Kearns M, Neel S, Roth A, Schutzman Z (2019) Fair algorithms for learning in allocation problems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp 170–179
Faber JW (2018) Segregation and the geography of creditworthiness: Racial inequality in a recovered mortgage market. Hous Policy Debate 28(2):215–247
Farnadi G, Babaki B, Getoor L (2018) Fairness in relational domains. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 108–114
Feldman M, Friedler SA, 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
FICO@score (2022) FICO score: the score lenders use
Fuster A, Plosser M, Schnabl P, Vickery J (2019) The role of technology in mortgage lending. The Review of Financial Studies 32(5):1854–1899
Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A (2022) Predictably unequal? the effects of machine learning on credit markets. J Financ 77(1):5–47
Ghosh A, Dutt R, Wilson C (2021) When fair ranking meets uncertain inference. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1033–1043
Giacoletti M, Heimer R, Yu EG (2021) Using high-frequency evaluations to estimate discrimination: Evidence from mortgage loan officers. In: Proceedings of Paris December 2021 Finance Meeting EUROFIDAI-ESSEC
Gogoll J, Müller JF (2017) Autonomous cars: in favor of a mandatory ethics setting. Sci Eng Ethics 23(3):681–700
Gordaliza P, Del Barrio E, Fabrice G, Loubes J-M (2019) Obtaining fairness using optimal transport theory. In: International conference on machine learning, pp 2357–2365
Grgić-Hlača N, Engel C, Gummadi KP (2019) Human decision making with machine assistance: An experiment on bailing and jailing. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–25
Grgic-Hlaca N, Zafar MB, Gummadi KP, Weller A (2016) The case for process fairness in learning: feature selection for fair decision making. In: NIPS symposium on machine learning and the law, vol. 1, p. 2. Barcelona, Spain
Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. Advances in neural information processing systems 29
Hassani BK (2021) Societal bias reinforcement through machine learning: a credit scoring perspective. AI and Ethics 1(3):239–247
Haupert T (2022) The racial landscape of fintech mortgage lending. Hous Policy Debate 32(2):337–368
Heckman JJ (1979) Sample selection bias as a specification error. Econometrica: Journal of the econometric society, 153–161
Hofmann H (1994) Statlog (german credit data) data set. UCI Repository of Machine Learning Databases 53
Hort M, Sarro F (2021) Did you do your homework? raising awareness on software fairness and discrimination. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp 1322–1326. IEEE
Hu L, Chen Y (2020) Fair classification and social welfare. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 535–545
Joseph M, Kearns M, Morgenstern JH, Roth A (2016) Fairness in learning: classic and contextual bandits. Adv Neural Inf Process Syst 29
Kallus N, Mao X, Zhou A (2022) Assessing algorithmic fairness with unobserved protected class using data combination. Manage Sci 68(3):1959–1981
Kallus N, Zhou A (2018) Residual unfairness in fair machine learning from prejudiced data. In: International Conference on Machine Learning, pp 2439–2448. PMLR
Kamiran F, Calders T (2012) Data preprocessing techniques for classification without discrimination. Knowl Inf Syst 33(1):1–33
Karimi A-H, Schölkopf B, Valera I (2021) Algorithmic recourse: from counterfactual explanations to interventions. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp 353–362
Kassam A, Marino P (2021) Algorithms, racism, and equity: a social impact approach. https://feminism-social-justice-ai.org/algorithms-racism-and-equity-a-social-impact-approach/
Kearns M (2017) Fair algorithms for machine learning. In: Proceedings of the 2017 ACM conference on economics and computation, pp 1–1
Kearns M, Roth A, Wu ZS (2017) Meritocratic fairness for cross-population selection. In: International conference on machine learning, pp 1828–1836
Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering-a systematic literature review. Inf Softw Technol 51(1):7–15
Kleinberg J, Mullainathan S, Raghavan M (2016) Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807
Knight E (2019) Ai and machine learning-based credit underwriting and adverse action under the ecoa. Bus. & Fin. L. Rev. 3:236
Kordzadeh N, Ghasemaghaei M (2022) Algorithmic bias: review, synthesis, and future research directions. Eur J Inf Syst 31(3):388–409

Kozodoi N, Jacob J, Lessmann S (2022) Fairness in credit scoring: Assessment, implementation and profit implications. Eur J Oper Res 297(3):1083–1094
Kumar A, Sharma S, Mahdavi M (2021) Machine learning (ml) technologies for digital credit scoring in rural finance: A literature review. Risks 9(11):192
Ladd HF (1998) Evidence on discrimination in mortgage lending. J Econ Perspect 12(2):41–62
Latour B (1986) Visualization and cognition. Knowledge and society 6(6):1–40
Le LH, Stefańczyk JK (2018) Gender discrimination in access to credit: are women-led smes rejected more than men-led? Gend Technol Dev 22(2):145–163
Lee MSA, Floridi L (2021) Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. Mind Mach 31(1):165–191
Lessmann S, Baesens B, Seow H-V, Thomas LC (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur J Oper Res 247(1):124–136
Li Y (2021) Gender differences in car loan access: An empirical analysis. In: The 2021 12th International Conference on E-business, Management and Economics, pp 493–498
Liu W, Guo J, Sonboli N, Burke R, Zhang S (2019) Personalized fairness-aware re-ranking for microlending. In: Proceedings of the 13th ACM Conference on Recommender Systems, pp 467–471
Liu Y, Radanovic G, Dimitrakakis C, Mandal D, Parkes DC (2017) Calibrated fairness in bandits. arXiv preprint arXiv:1707.01875
Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R (2019) Bias mitigation post-processing for individual and group fairness. In: Icassp 2019-2019 Ieee International Conference on Acoustics, Speech and Signal Processing (icassp), pp 2847–2851. IEEE
Loya J (2022) Racial stratification among latinos in the mortgage market. Race Soc Probl 14(1):39–52
Maaitah NA (2018) Discriminatory practice in microfinance: Gender and glass ceiling on loan size (case study from jordan). Journal of Central European Green Innovation 6(1063-2018-4223), 35–54
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR) 54(6):1–35
Mehrabian A, Russell JA (1974) An Approach to Environmental Psychology. the MIT Press, Cambridge
Mendes LS, Mattiuzzo M (2022) Algorithms and discrimination: the case of credit scoring in brazil. In: Albers M, Sarlet IW (eds) Personality and Data Protection Rights on the Internet, vol 96. Springer, Switzerland, pp 407–443
Miconi T (2017) The impossibility of "fairness": a generalized impossibility result for decisions. arXiv preprint arXiv:1707.01195
Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2021) Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8:141–163
Mitchell S, Shadlen J (2017) Fairness: Notation, definitions, data, legality
Moscato V, Picariello A, Sperlí G (2021) A benchmark of machine learning approaches for credit score prediction. Expert Syst Appl 165:113986
Nyarko SA (2022) Gender discrimination and lending to women: The moderating effect of an international founder. Int Bus Rev 31(4):101973
Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453
Otieno B, Wabwoba F, Musumba G (2020) Towards small-scale farmers fair credit scoring technique. In: 2020 IST-Africa Conference (IST-Africa), pp 1–11. IEEE
Park KA (2022) A comparison of mortgage denial and default rates by race, ethnicity, and gender. Ethnicity, and Gender (February 7, 2022)
Phelps ES (1972) The statistical theory of racism and sexism. Am Econ Rev 62(4):659–661
Pi T, Liu Y, Song J (2020) Does geographical discrimination exist in online lending in china: An empirical study based on chinese loan platform renren. International Journal of Financial Studies 8(1):15
Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ (2017) On fairness and calibration. Adv Neural Inf Process Syst 30
Prasad P (2022) True colors of global economy: In the shadows of racialized capitalism. Organization, 13505084211066803
Ragnedda M (2020) New digital inequalities. algorithms divide. In: Enhancing Digital Equity, pp 61–83. Springer, Switzerland
Rebitschek FG, Gigerenzer G, Wagner GG (2021) People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. Sci Rep 11(1):1–11
Ruparelia NB (2010) Software development lifecycle models. ACM SIGSOFT Softw Eng Notes 35(3):8–13
Russell C, Kusner MJ, Loftus J, Silva R (2017) When worlds collide: integrating different counterfactual assumptions in fairness. Adv Neural Inf Process Syst 30
Sackey FG, Amponsah PN (2020) Information asymmetry and self denial in gender participation in commercial banks' credit markets in emerging economies in ghana. Journal of Small Business & Entrepreneurship, 1–28
Sackey FG, Amponsah PN (2018) Gender discrimination in commercial banks' credit markets in ghana: a decomposition and counterfactual analysis. African Journal of Business and Economic Research 13(2):121–140
Salgado CCR, Aires RFdF (2018) Microcredit and gender: Are there differences in the credit conditions? BAR-Brazilian Administration Review 15
Salimi B, Rodriguez L, Howe B, Suciu D (2019) Interventional fairness: Causal database repair for algorithmic fairness. In: Proceedings of the 2019 International Conference on Management of Data, pp 793–810
Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y (2019) How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 99–106
Schoeffer J, Kuehl N (2021) Appropriate fairness perceptions? on the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In: Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing, pp 153–157
Schoeffer J, Kuehl N, Valera I (2021) A ranking approach to fair classification. In: ACM SIGCAS Conference on Computing and Sustainable Societies, pp 115–125
Segal S, Adi Y, Pinkas B, Baum C, Ganesh C, Keshet J (2021) Fairness in the eyes of the data: Certifying machine-learning models. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp 926–935
Singh A, Singh J, Khan A, Gupta A (2022) Developing a novel fair-loan classifier through a multi-sensitive debiasing pipeline: Dualfair. Mach Learn Knowl Extract 4(1):240–253
Steil JP, Albright L, Rugh JS, Massey DS (2018) The social structure of mortgage discrimination. Hous Stud 33(5):759–776
Stigler GJ (1987) The theory of price. Macmillan, New York
Stiglitz JE, Weiss A (1992) Asymmetric information in credit markets and its implications for macro-economics. Oxf Econ Pap 44(4):694–724
Streltfeld D (2020) On the Web, Price Tags Blur. https://www.washingtonpost.com/archive/politics/2000/09/27/

on-the-web-price-tags-blur/14daea51-3a64-488f-8e6b-c1a3654773da/
Sun H, Gao L (2019) Lending practices to same-sex borrowers. Proc Natl Acad Sci 116(19):9293–9302
Swan SL (2019) Discriminatory dualism. Ga. L. Rev. 54:869
Tran TKV, Elahi E, Zhang L, Abid M, Pham QT, Tran TD (2018) Gender differences in formal credit approaches: rural households in vietnam. Asian-Pacific Economic Literature 32(1):131–138
Valentim I, Lourenço N, Antunes N (2019) The impact of data preparation on the fairness of software systems. In: 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE), pp 391–401. IEEE
Wakchaure MA, Sane SS (2018) An algorithm for discrimination prevention in data mining: Implementation statistics and analysis. In: 2018 International Conference On Advances in Communication and Computing Technology (ICACCT), pp 403–409. IEEE
Wang S, Gupta M (2020) Deontological ethics by monotonicity shape constraints. In: International conference on artificial intelligence and statistics, pp 2043–2054
Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering, pp 1–10
Wong P-H (2020) Democratizing algorithmic fairness. Philosophy & Technology 33(2):225–244
Yu E (2022) Banking trends discrimination in mortgage markets. Banking Trends 7(1):2–8
Zafar MB, Valera I, Rogriguez MG, Gummadi KP (2017) Fairness constraints: mechanisms for fair classification. In: Artificial intelligence and statistics, pp 962–970
Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp 335–340