Uncovering the Source of Evaluation Bias in Micro-Lending

Completed Research Paper

Xiyang Hu*
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
xiyanghu@cmu.edu

Yan Huang*
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
yanhuang@cmu.edu

Beibei Li*
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
beibeili@andrew.cmu.edu

Tian Lu*
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
lutiansteven@gmail.com

Abstract

We develop a structural econometric model to capture the decision dynamics of human evaluators on an online micro-lending platform, and estimate the model parameters using real-world data. We find two types of biases in gender, i.e. preference-based bias and belief-based bias, are present in human evaluators' decisions. Both types of biases are in favor of female applicants. Through counterfactual simulations, we quantify the effect of gender bias on loan granting outcomes and the welfare of the company and the borrowers. Our results imply that both the existence of the preference-based bias and that of the belief-based bias reduce the company's profits. When the preference-based bias is removed, the company earns more profits. When the belief-based bias is removed, the company's profits also increase. Both increases result from lowering the approval probability for borrowers, especially female borrowers, who eventually default on loans. For borrowers, the elimination of either bias decreases the gender gap in the credit risk evaluation.

Keywords: Bias, Discrimination, Structural Modeling, Dynamic Behavior, Micro-loan

** Author names are in alphabetical order.*

Introduction

Humans and organizations often need to make decisions under imperfect information, and they generally rely on certain statistics that quantify the likelihood of different outcomes (Corbett-Davies and Goel 2018). However, in practice these statistics generally cannot perfectly predict the outcome of the event, and often involve human inputs, which may contain bias against certain demographics (often referred to as protected groups) such as minorities or females. As a result, members in these demographic groups are often treated unfairly, which leads to various social and economic problems.

A typical context of decision making under imperfect information is the loan approval decision in micro-lending. The emerging micro-lending business provides faster and convenient access to financial resources to more people with a streamlined loan application process (Mateescu 2015). The application process does not demand visits to a physical institution, and the credit assessment process involves human decisions (Lu et al. 2012). That is, human evaluators (i.e. platform staffs or lenders) make credit risk assessment with individual information (e.g., demographics) which is collected from loan applicants. Based on the evaluated credit risk, evaluators make the final loan approval decisions. However, this practice may contain bias due to limited human cognitive capacity to process complex computation for credit risk evaluation (Icard 2018). Theoretically, bias in credit risk evaluation would heavily attenuate the evaluation accuracy, profitability of micro-lending platform or lenders, and cause social unfairness (Fu et al. 2019). Unfortunately, the influence of human bias is largely neglected in literature. Since human evaluations are inevitable in many contexts such as micro-lending, the goal of this study is to investigate: **Does bias exist in human evaluators' loan evaluation; if so, how does this bias affect their decisions and the market outcome?**

Particularly, over the past decades, economists and social scientists have proposed different models to explain and quantify human bias (Bohren et al. 2019, Arnold et al. 2018, Gneezy et al. 2012, Parsons et al. 2011, Bertrand et al. 2005, Biernat and Manis 1994). They classify human bias into two broad categories: *preference-based bias* (or tasted-based bias) and *belief-based bias* (Bohren et al. 2019). Preference-based bias arises when evaluators have animus towards a particular group, while belief-based bias arises when evaluators' subjective beliefs about a group lead them to less favorably treat individuals from the group than members from other groups with the same observed performance. The classification of the two types of human bias is critical. It allows us to dynamically trace the evolution of bias during long-term (i.e., repeated) decisions rather than to regard bias as static. This static assumption of bias may overshadow the potentials and value of learning behaviors in decision making (Zhang and Angela 2013). It also enables us to unravel the influence of different types of human bias on decision outcomes.

Therefore, following such a classification of the source of the decision bias, in this study, we disentangle and quantify these two types of human bias in observable data from a micro-lending platform. In specific, we develop a structural econometric model of human evaluators' loan approval decisions, which captures the underlying economic processes that drive human evaluators' decisions. We then estimate the structural model based on the real-world data from the micro-lending platform that involves sufficient samples of repeated loan applications and approval decisions by the platform staff (i.e., human evaluators). Some of the parameters in the model capture the two possible types of human biases, and therefore, their parameter estimates can reveal whether human evaluators' decisions exhibit those two types of bias. We compare a number of alternative model specifications and identify the one that best explain the observed human decisions in the data. Using the estimated structural model, we conduct policy simulations to quantify the effects of these human bias and their welfare implications.

In the counterfactual simulations, we take advantage of a unique experiment conducted on the platform, that is, during a period of time, all applicants are approved without any selection. This novel experiment allows us to observe the credit behaviors of all the applicants on the micro-lending platform which are usually unobservable. This allows us to infer the outcomes of the loans that are not approved by human evaluators, which usually go unobserved and are impossible to evaluate.

Specifically, to measure the extent of bias, we examine the gender gaps for the approval true positive rate (TPR). We adopt the concept of *equal opportunity*, one of the most popular fairness notions. *Equal opportunity* requires that qualified individuals, no matter what their sensitive attributes are, have equal opportunity to receive favorable outcomes (Hardt et al. 2016). In our loan application setting, this means two gender group should have the same true positive rates where the positive label is "non-default".

We find that the evaluators exhibit both types of biases discussed above, with both preference-based bias and belief-based bias in favor of females. Although preference-based bias persists, belief-based bias gradually reduces as human evaluators learn from the repayment behaviors of each specific borrower. Our policy simulations suggest that the elimination of either the preference-based bias or the belief-based bias from human evaluators' decisions can increase the platform's profits. The mechanisms behind the two scenarios are the same. The extra profits result from lowering the approval probability of defaulters, especially female defaulters. On the borrower end, the elimination of the two types of bias can mitigate the gender gap in the credit evaluation measured by the true positive rate. And when both types of bias are removed at the same time, the gender gap is minimized.

The theoretical contribution of this study is multi-fold. First, we introduce two types of human biases into a real-world decision-making context, and we are among the first to empirically identify and quantify these two types of human biases with a large-scale dataset and a structural econometric model. Second, we also add to the micro-lending and FinTech literature by revealing how human bias could influence platforms' profitability and service equality. We also characterize how human evaluators make credit risk evaluations and loan approval decisions in practice. Methodologically, we design a novel and comprehensive empirical framework to uncover the behavioral sources of bias, which are usually implicit and difficult to identify accurately. The framework works on secondary datasets and enables to conduct counterfactual simulations.

Relevant Literature

Our work is closely related to the literature on the bias and discrimination in human decision making, especially bias in financial evaluations. Two categories of human decision biases have been extensively studied; one is ethnic or racial discrimination, and another is gender discrimination. It has been documented that in financial loan market, both non-Fintech and Fintech lenders tend to discriminate against ethnic-minority borrowers (higher interest rates and lower probability of being funded) through liability document and facial bias (Sydnor and Pope 2011, Bartlett et al. 2019). It has also been shown that women are more credit constrained than men by microfinance institutions (Blanchflower et al. 2003). In P2P lending, female borrowers need pay higher interest rates (Alesina et al. 2013, Chen et al. 2017, 2020). And female founders are less successful attracting male investors compared to observably similar male founders (Ewens and Townsend 2020).

In addition to race bias and gender bias, other kinds of bias or discrimination common to see in society include immigration and age related bias (Dobbie et al. 2018), occupational related bias (Cui 2019), home bias (Lin and Viswanathan 2016), etc. The major reasons behind these human decision bias lie in the minority applicants' relative quality compared to the majority (Ferguson and Peters 1995), the decision makers' improper task objectives and incentives (Dobbie et al. 2018), and the inherent bias formation and evolution process of preference-based and belief-based evaluation biases (Gneezy et al. 2012, Bohren et al. 2019).

Economic theories classify inherent human bias into two types: preference-based bias and belief-based bias (Bohren et al. 2019). Preference-based bias arises when evaluators have animus towards a particular group, while belief-based bias arises when evaluators' subjective beliefs about a group lead them to less favorably treat individuals from the group than members from the regular group with the same observed performance. Belief-based bias can be further classified into two subcategories: belief-based bias with misspecified/incorrect beliefs and belief-based bias with correct beliefs (sometimes referred to as statistical bias). The former occurs when the evaluators' subjective beliefs about the group-level statistics of the protected group are not the same as the reality, and the latter occurs when the subjective beliefs match the reality. In a static setting, preference and belief-based biases mix up with each other, and it is hard to disentangle their effects on human decisions. However, in a dynamic setting, we are able to distinguish between the two biases. This is because across periods, preference-based bias is likely to persist, while belief-based bias can be mitigated or even reversed when the evaluator observes new signals about each individual they are evaluating. In this paper, we follow the definitions of the two types of bias. Our context of multi-period microloan borrowing and lending provides a great setting to identify these two types of bias.

Our paper also builds on the growing literature on algorithmic discrimination and machine learning bias. One important source of machine bias is the human decision bias encoded in the training dataset. Fuster et al. (2020) incorporate predictions of machine learning models into a simple equilibrium model of financial credit, and find that algorithms increase rate disparity among and within different racial groups. Stevenson and Doleac (2019) conduct simulations to evaluate the race and age disparities in finance risk assessment, and demonstrate that human and machine interaction can lead to bias in both race and age. Lambrecht and Tucker (2019) find that economic forces in the market can distort neutral algorithms into discriminating females in terms of their exposure to advertisements of STEM (science, technology, engineering and mathematics) jobs. Major reasons for algorithmic bias include lack of necessary data control (statistical bias) and unintended correlation with sensitive factors (Fu et al. 2019, Bartlett et al. 2019), training-sample bias (Cowgill et al. 2020), market mechanism (Lambrecht and Tucker 2019), etc. Even though algorithms could lead to various bias issues, appropriate designs and regulations can make them positive forces for equity (Kleinberg et al. 2018b, Chouldechova 2017, Rudin 2019).

To deal with bias problems in human and machine decision making, researchers have come up with diverse methods. The most direct way is to obtain and include more useful data. Kleinberg et al. (2018a) show direct inclusion of a protected variable (e.g., race) is useful for mitigating unfairness. Lu et al. (2019) find that using proper alternative data could improve both financial profitability and equality. Aside from enriching the data, learning from repeated events can also mitigate bias. Cai et al. (2016) conclude that based on signaling theory, evaluators/investors will leverage information from repeated borrowing of the same borrower in her lending history. Kim (2020) argues that borrowers' past track record within the platform have the most important impact (than other demographic factors) on predicting the repayment performance of their current loans.

Another flourishing track to combat decision bias is to invent more transparent, delicate and de-biasing algorithms. It has been widely shown that enhanced algorithmic transparency and interpretability can help eliminate bias, and sometimes simple and transparent models are able to outperform complex black-box models (Rudin et al. 2020, Rudin and Shaposhnik 2019, Rudin 2019, Hu et al. 2019). Wang et al. (2013) demonstrate that Bayesian investment model can significantly improve investors' investment decisions based on other investment models. Choudhury et al. (2020) argue that human capital and machine learning can complement each other through combining algorithms and domain expertise or knowledge. Many other de-biasing methods draw from the perspective of statistics, optimizations, and behaviors, etc (Berk et al. 2017, Lum and Johndrow 2016, Hardt et al. 2016, Kamiran et al. 2010, Fu et al. 2021).

In terms of decision quality, human predictions often tend to be less accurate, which can negatively affect the quality of their decisions. This is because on the one hand, people may have resource-limited brain to process complex computation in evaluation (Icard 2018); on the other hand, people may use a simple updating rule, which, for example, linearly combines their personal experience and accumulated knowledge for repeated tasks (Jadbabaie et al. 2012).

Compared with human decision making, algorithm-based decision making has demonstrated superior ability to achieve better accuracy and handle more complex information. Human v.s. machine decision making has been widely studied in healthcare area. In most cases, machine learning models outperformed or tied the judgment accuracy of an average clinician (Camerer 2019), and only a small fraction of clinicians were more accurate than machine learning models (Goldberg 1970, et al.). Mechanical-prediction techniques were about 10% more accurate than clinical predictions (Grove et al. 2000). Very simple actuarial methods (i.e., linear combination of criterion variables) has been shown to consistently perform better than clinical judgment (Dawes 1971).

However, there are also scenarios in which human experts can outperform machines. Some examples are tasks that heavily require theory-driven judgement that are not suitable for statistical models; rare events or outliers that have never been seen by algorithms; complex configural relationships between the features and the dependent variable (Dawes et al. 1989). In our context, micro-loan approval decisions do not face these problems that make humans better than machines. And in finance industry, algorithms are widely used to identify credit risks, with XGBoost (Chen and Guestrin 2016) as the most popular one.

Methodologically, our paper also builds upon the abundant work on modeling human decision dynamics through structural models. Erdem et al. (2008) use a Bayesian learning framework to model consumers' brand choices under quality signals from advertisement, price and past consumption experiences. Huang et al. (2014) investigate the learning dynamics of users' idea posting behavior on a crowdsourcing platform. Zhang et al. (2019) study participants' learning behavior from superstars in crowdsourcing contests. Zhang et al. (2020) examine taxi drivers' learning behavior based on fine-grained GPS observations. In this paper, we model the learning dynamics in loan application evaluators' decisions, and disentangle and estimate the preference-based bias and belief-based bias in their behavior.

Research Context and Data

Context

We obtained our data from a leading Asian micro-lending platform. The platform was founded in 2011 and offers microloans at an average size of approximately 450 USD. Loan applicants on the platform use the loans primarily to fulfill temporary financial needs including supplementary working capital for small businesses, irregular shopping needs, education spending, and medical expenses. To apply for a loan, applicants must provide their personal information such as name, gender, age, income level, and a copy of their national identity cards. They must also provide their contact persons. People under the age of 18 and all students at school or university are not qualified to apply as they usually have no independent income sources. The loan term ranges from one to eight months. The annual interest rate charged by the platform is approximately 18% (plus or minus 1%, depending on the credit line of the borrowers).

During our sample period, the platform evaluates applicants' credit risk manually by its employees (i.e., evaluators). All the evaluators are trained regularly to maintain consistent evaluation criteria, which are derived from their collective daily work experience. Besides, no gender bias training has been conducted by the platform. They do not use any AI technologies (e.g., machine learning) as automatic or auxiliary tools for evaluation. In specific, after an applicant fills in her personal information and submits the application in the system, she will be randomly assigned to an evaluator. Then, the evaluator decides whether to issue the loan by assessing whether the applicant can bring positive economic benefit (i.e., profit) to the platform. Loan profit is roughly calculated based on the predicted probabilities of delinquency and default. If a borrower fails to repay an installment, she will be regarded as delinquent. If a loan is unpaid 90 days or more after the due date, default is confirmed by the platform.

The estimations of the chances of delinquency and default are based on the collected personal information for all new applicants. For repeated applicants, evaluators will additionally leverage information from the applicants' repayment performance on previous loans, such as the final overdue days (denoted as D), the proportion of overdue installments (denoted as M), the proportion of installments with positive attitude from the borrower (denoted as A), which is measured from the records of whether a borrower has shown a positive attitude towards their financial obligation during her communication with the platform, and the proportion of installments with financial help from family or friends (denoted as H).

When a borrower becomes delinquent, the platform will impose financial penalty on them. Simultaneously, debt collection methods such as sending reminder notifications to them and their contact persons will be implemented. Borrowers in default are prohibited from applying for loans again on the focal platform. Default records are also submitted to the personal credit record system maintained by the central government and a shared blacklist system maintained by a symposium of micro-finance institutions. The platform may take legal actions against defaulters.

Data and Description

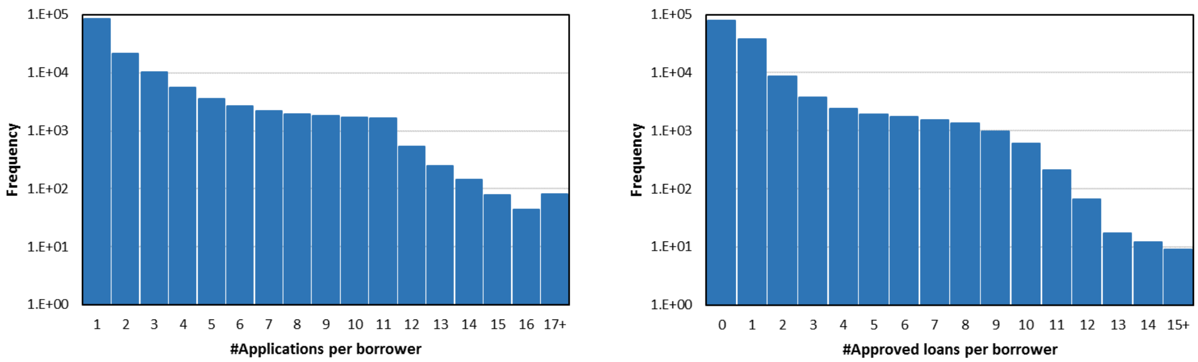
Our data set contains fine-grained information of both the applicants whose submitted applications were approved and those whose applications were rejected by the platform between January 2015 and September 2017 (i.e., 33 months). During the sample period, there are 311,200 loan applications in total, among which 135,938 loan applications (taking up or approval rate 43.68%) were approved, whereas 175,263 were rejected by the platform. Our sample covers 139,454 borrowers; that is, the average number (frequency) of loan applications per borrower is 2.23 (= 311,200/139,454). In our sample, 53,503 (38.37%) borrowers applied more than once, and they contributed 225,248 applications in total (i.e., 4.21 on average per borrower). For these multiple-time borrowers, the average approval rate is 47.24%. The average approval rate is only 34.34% for the 85,951 borrowers who applied just once. This indicates the platform's preference towards repeated borrowers, which is reasonable as these borrowers have performed well in historical loans. Figures 1a and 1b display the distributions of the frequency of loan applications and number of approved loans respectively.

		Observations	Repeated applicants		New applicants	
			Mean	S. D.	Mean	S. D.
Applicant information (Female/Male)	Gender (1 = female)	53,503/85,952	0.19	0.39	0.18	0.39
	Education level	53,503/85,952	2.25	0.64	2.27	0.63
	Monthly income level	53,503/85,952	3.27	1.80	3.37	1.90
	Home city DPI	53,503/85,952	2.34	1.31	2.54	1.49
	House ownership (1= self-own)	53,503/85,952	0.17	0.38	0.16	0.37
Loan information	Loan amount (USD)	311,200	460.70	81.59	458.78	321.30
	Loan term (month)	311,200	5.88	1.54	5.86	1.35
	Yearly interest rate (%)	311,200	14.05	1.28	14.36	1.44

Note. *Education level: 1 = middle school; 2 = technical school; 3 = undergraduate; 4 = postgraduate.*

Table 1. Information of the Approved and Rejected Applications

For all the applicants, we obtain their demographic and socioeconomic data including gender, education level, monthly income level, disposable personal income per capita (DPI) of their home city, and house ownership, and their loan information including loan amount, loan term, and annual interest rate. We likewise have detailed per-installment repayment information of the approved loans. Table 1 summaries information of the approved and rejected applications.



(a) Frequency of loan applications (b) Number of approved loans

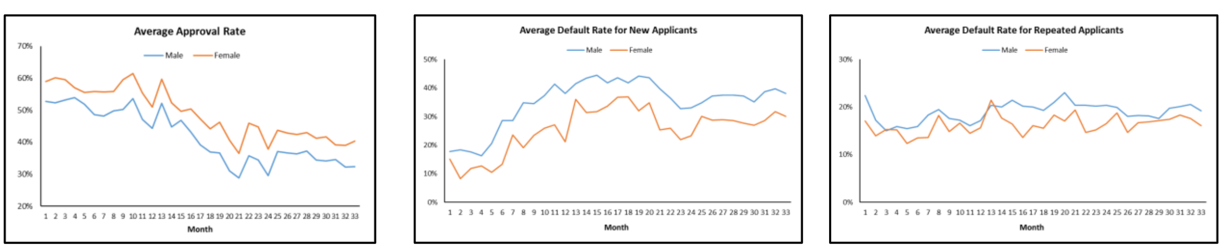
Figure 1. Statistics of Loan Applications

Model-Free Analyses

On the platform, female applicants are generally more likely to be approved. The gaps between the female and male approval rates do not shrink (Figure 2a) over time, implying that the platform evaluators may have a persistent impression of credit risks between males and females.

When we consider only borrowers who have applied repeatedly (Figure 3), the approval gender gap shrinks, suggesting that learning helps amend the platform evaluators' prior bias in gender. Figure 4 shows the trend of the default rate as borrowers as the number of applications or the number of their previously approved loans increases. Consistently, the gender gap in the default rate is smaller for repeated borrowers; and it keeps decreasing as the borrowers' number of previous applications increases, and as number of previous approvals increases. This is consistent with the decreasing gender gap in the approval rate, which indicates that evaluators can learn effectively from users' previous repayment behaviors and adjust their prior belief-based bias in gender.

As noted earlier, the evaluator relies on four signals from the users' past repayment behaviors to make approval decisions on loan applications. Figure 5 shows the values of the four signals against the number of previous applications by genders. Generally, we find that, given the number of previous applications, all the four signals show similar values between females and males, indicating that these signals may help the platform evaluators to adjust their prior bias in gender.



(a) Average default rate for all applicants (b) Average default rate for new applicants (c) Average default rate for repeated applicants

Figure 2. Time Trends of Approval and Default Rate

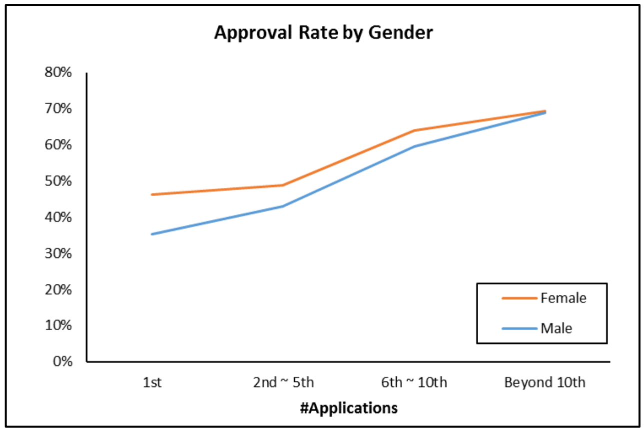
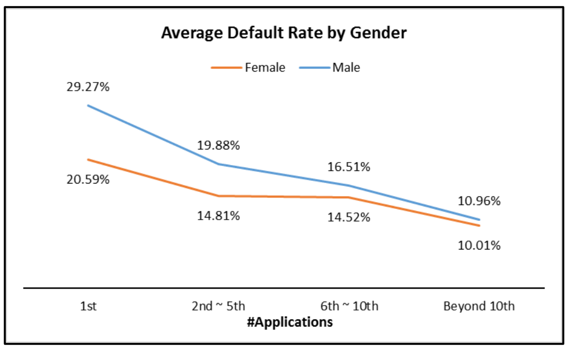
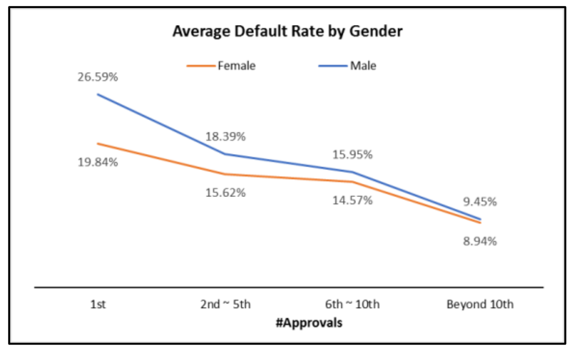


Figure 3. Approval Rate by Gender



(a) Average Default Rate by Gender for users with different number of applications



(b) Average Default Rate by Gender for users with different number of approvals

Figure 4. Average Default Rate by Gender

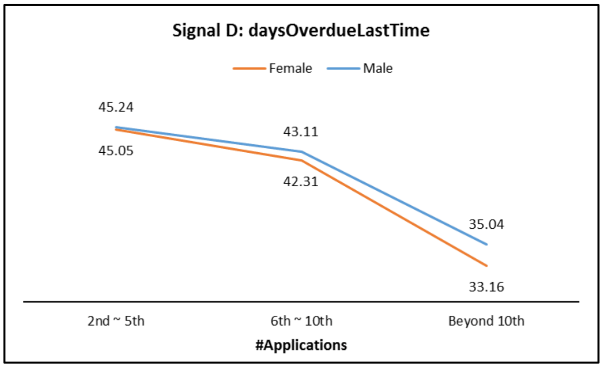
Model

We consider a model in which the loan platform decides whether or not to approve a loan application by applicant i at time t (here t indicates the t-th application rather than a natural time unit). We model the evaluator behavior in a dynamic environment where the evaluator is uncertain about the true credit quality of applicants. When a new borrower comes to the micro-lending platform, the evaluator only observes her demographics and form a prior belief based on the demographic data. For every borrower, without any previous repayment behavior being observed, her first application is preprocessed by the evaluator using the prior belief. If a borrower's loan application is approved at t, then at the time of her next loan application, i.e., t + 1, the evaluator will use the previous repayment behaviors as additional signals of the borrower's credit quality.

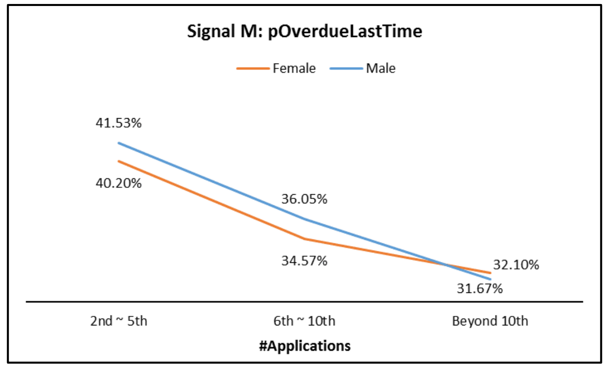
At time $t = 0$, without any observation of borrower i's repayment behaviors, given i's demographics X_i, the evaluator forms a belief of her credit quality with mean βX_i and variance $\sigma^2_{Q_0}$. X_i includes *gender, age, education level, marriage status, house ownership, monthly income, the disposable personal income (DPI) of borrowers' living cities (in 2017)*. We also incorporate a constant 1 and *the time of* i*'s first application* into X_i, because the overall borrowers' qualities may keep changing over time. We assume that the prior belief follows a normal distribution:

$$Q_{i0} \sim N\left(\beta X_i, \sigma^2_{Q_0}\right) \text{ for i} = 1, \dots, \text{N}. \tag{1}$$

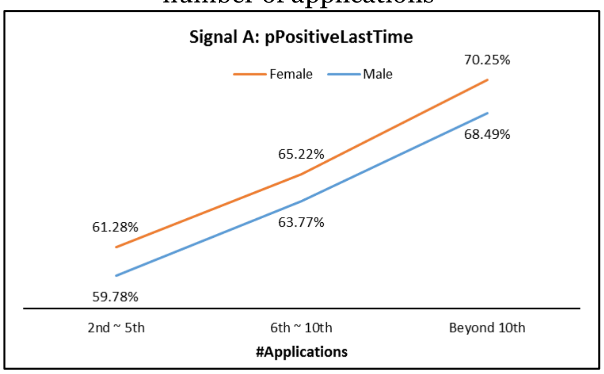
where βX_i is the evaluator's subject prior belief about the mean credit quality of a borrower with demographics X_i. The coefficient for gender is β_g, where gender $g \in \{M, F\}$, M stands for males and F stands for females. We normalize β_M to be zero. Therefore β_F captures the **belief-based bias**.



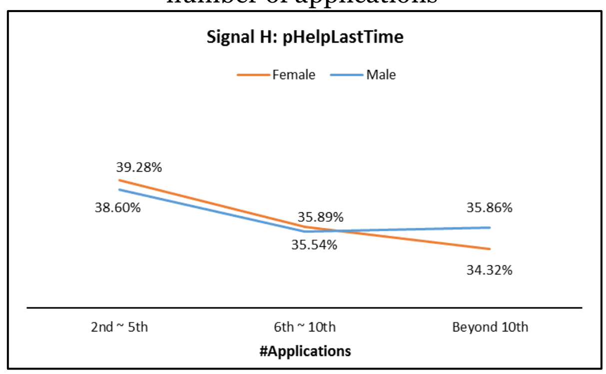
(a) Signal D by Gender for users with different number of applications



(b) Signal M by Gender for users with different number of applications



(c) Signal A by Gender for users with different number of applications



(d) Signal H by Gender for users with different number of applications

Figure 5. Signals Across Number of Applications

For every borrower, we assume that the evaluator believes that their credit qualities are of identical uncertainty, i.e., $\sigma^2_{Q_0}$ is identical for all i.

When the evaluator processes the loan application of borrower i at time t, she has an information set I_{it}, which contains all loan repayment histories up to the loan of time $t-1$. Given the information set, the evaluator form a posterior belief of Q_i. Let Q_{it} be the expectation of the evaluator's belief for i's credit quality at time t, i.e.

$$Q_{it} = \mathrm{E}[Q_i | I_{it}] \tag{2}$$

At time t, when the evaluator is reviewing the loan application from borrower i, she observes four signals from i's repayment behavior on previous loans.

Consider when an applicant i applies for a loan at time t, the evaluator observes four signals of the repayment behavior of her last loan, the final overdue days D_{it}, the proportion of overdue installments M_{it}, the proportion of installments with positive attitude from the borrower A_{it}, and the proportion of installments with financial help from family or friends H_{it}. Some of these signals may be more informative than others. Therefore, we tried several different combinations of these signals, and drop M_{it} from the signal list in our final model as M_{it} turns out to be uninformative and does not affect evaluators' loan approval decisions significantly.

In previous literature (Erdem et al. 2008, Huang et al. 2014, Zhang et al. 2020, et al.), it is common to model such behaviors using Bayesian updates of a normal-normal conjugate prior-posterior. We also start with such a conventional normal-normal conjugate Bayesian model, i.e., each time when additional signals become available, the evaluator updates her belief of a borrower's credit quality in a Bayesian fashion. However, the estimated parameters of such a Bayesian learning model show that all variances of these signals are extremely small. This indicates that evaluators weight the most recent signals heavily and are not updating their beliefs in a Bayesian fashion. Therefore, we use a simplified model to capture the evaluators' updating behaviors, where the posterior belief is a weighed sum of the prior belief and the new signals. We tried different combinations of the four signals. The weighted sum model with signals D_{it}, A_{it}, and H_{it} performs the best and achieves the highest likelihood. We use this model as our main model.

Because the final overdue days D_{it} has a long-tail distribution, we take the logarithm of D_{it} and use $\log D_{it}$ for subsequent calculations. We assume the evaluators believe that all these signals are linearly related to the credit quality in the following way:

$$\log D_i^M = D_0 + \phi Q_i,$$
$$A_i^M = A_0 + \psi Q_i,$$
$$H_i^M = H_0 + \rho Q_i,$$
$$(3)$$

where ϕ, ψ, and ρ are parameters of slopes, D_0, A_0, and H_0 are corresponding intercepts, and $\log D_i^M$, A_i^M, and H_i^M are the means of D_{it}, A_{it}, and H_{it}. We assume each borrower's signals are distributed surrounding their means.

At each time t when the evaluator observes these repayment behaviors, she updates her belief about the credit quality based on a weighted sum of the prior and the signals:

$$Q_{it} = (1 - \alpha_D - \alpha_A - \alpha_H)Q_{i,t-1} + \alpha_D \frac{\log D_{it} - D_0}{\phi} + \alpha_A \frac{A_{it} - A_0}{\psi} + \alpha_H \frac{H_{it} - H_0}{\rho} \tag{4}$$

where α_D, α_A, and α_H are the weights assigned to the three signals. At time t, the evaluator decides whether to approve i's loan application based on her updated belief of the applicant's quality Q_{it}. The evaluator first calculates the probability of non-default through a sigmoid function:

$$p_{it} = h(Q_{it}) = \frac{1}{1 + \exp(-Q_{it})} \tag{5}$$

Then, with the probability of non-default p_{it}, the evaluator decides to approve or reject i's loan application of time t through a utility function. There are two key components in the utility function (Equation 6). The first component captures the expected profit of approving this loan. The second component (c_{ig}) contains the evaluator's preference-based bias. And we also assume there is a random shock within the utility function. The utility function is as follows:

$$u_{it} = z * (p_{it}a_{it} - (1 - p_{it})b_{it}) - c_{ig} + \epsilon_{it}, \tag{6}$$

where $g \in \{M, F\}$, and M stands for males and F stands for females; c_{ig} is the preference-based bias with c_{iM} normalized to be zero. Therefore, c_{iF} captures the **preference-based bias**, which persists and is not affected by observing new signals. a_{it} is the profit earned by the platform if the loan is paid back, and b_{it} is the loss the platform incurs if the loan defaults. Both a_{it} and b_{it} are observed values in our dataset. z is the price parameter (or marginal utility of money). p_{it} is the non-default probability of i at time t, which is related to its quality Q_{it}. All these parameters are estimated through maximizing the likelihood.

Estimation Results

We report parameter estimates for our model in Table 2. Both the preference-based bias c_{iF} (we normalize c_{iM} to be zero) and the belief-based bias β_F have expected signs (c_{iF} =-0.2551, β_F =0.1133), implying that the evaluator has a preference for female loan applicants. The estimate of the belief-based bias β_F is significantly positive, suggesting that evaluators have a higher prior belief for females' credit qualities. The estimate of c_{iF} is significantly negative. This implies that there is a significant preference-based bias in gender that favors female applicants, which cannot be corrected by observing repayment behaviors.

Apart from β_F, all other βs also have expected signs. This is consistent with the evaluators' preference for applicants with a better socioeconomic status as we observe in the data.

Our estimates of the slopes in the signal D-quality, signal A-quality and signal H-quality relationships are negative ($\phi = -0.0239$), positive ($\varphi = 0.3563$) and positive ($\rho = 0.9741$) respectively. These results suggest that in the evaluator's decision-making process, larger final overdue days D_{it} are associated with poor credit quality; while the proportion of installments with a positive attitude A_{it} is positively associated with loan approvals. Getting financial help from family and friends H_{it} is also viewed by evaluators as a positive signal for credit quality and is associated with a lower default probability.

As can be seen in Equation 4, the evaluator updates her belief based on a weighted sum of the prior belief and the signals. The estimates of the weights α_D =0.0166, α_A =0.9689 and α_H =0.0144 indicate that the evaluator gives most weight to the signal A_{it}, with a weight of about 0.9689.

Parameters	Estimate	Std. error
c_{iF}	-0.2551***	0.0308
Signal D		
ϕ	-0.0239***	0.0007
D_0	0.5116***	0.0365
Signal A		
ψ	0.3563***	0.0030
A_0	0.3993***	0.0096
Signal H		
ρ	0.9741***	0.0181
H_0	0.1261***	0.0179
z	0.0163***	0.0001
Coefficients of the prior β		
β_0	-1.029***	0.0109
β_F	0.1133***	0.0119
$\beta_{firstAppMonth}$	-0.0199***	0.0003
$\beta_{housing}$	0.1318***	0.0055
$\beta_{education}$	0.2516***	0.0036
β_{income}	0.0829***	0.0012
β_{DPI}	0.1128***	0.0017
α_D	0.0166***	0.0006
α_A	0.9689***	0.0049
α_H	0.0144***	0.0005

Note: *p <0:1; **p <0:05; ***p <0:01

Table 2. Structural Model Estimation Results

In Table 3, we compare the characteristics of the actual observed approved users and the expected values of the characteristics of the approved users our structural model predicts. All these statistics are very similar between the actual observations and our model's predictions. This suggests our model captures the decision process well.

t	number of users		number of females		mean of housing		Mean of DPI		mean of education		mean of income	
	SM	actual	SM	actual	SM	actual	SM	actual	SM	actual	SM	actual
1	53539.33	51019	12113.01	11797	0.085023	0.082192	1.078531	1.038077	0.967054	0.92623	1.525907	1.470026
2	22371.7	22486	4742.72	4690	0.03179	0.031874	0.462848	0.464454	0.38965	0.39123	0.62198	0.626156
3	14110.73	13993	2802.512	2734	0.018555	0.018551	0.277881	0.27404	0.240154	0.23778	0.37979	0.376518
4	10139.22	10392	1967.239	2015	0.013	0.013295	0.192308	0.194846	0.17002	0.17449	0.26495	0.270864
5	7917.322	8335	1520.431	1605	0.009977	0.010598	0.146994	0.153377	0.1317	0.138656	0.202175	0.211797
6	6376.219	6915	1222.235	1314	0.008252	0.009007	0.117074	0.125568	0.105735	0.11472	0.161333	0.174509
7	5035.121	5481	970.1422	1056	0.006301	0.006884	0.092232	0.099043	0.08350	0.09083	0.126435	0.137056
8	3629.12	3968	692.8316	755	0.004384	0.004905	0.065735	0.0714	0.06027	0.06569	0.09011	0.098269
9	2342.886	2564	441.4362	484	0.002798	0.003091	0.042753	0.046345	0.038917	0.0426	0.05755	0.062881
10	1150.928	1272	219.6785	234	0.001276	0.001413	0.020986	0.02309	0.019207	0.02123	0.02936	0.032297
11	436.4425	475	80.60686	89	0.00049	0.000552	0.00845	0.009093	0.00725	0.00787	0.011423	0.012348
12	130.8604	129	21.06548	20	0.000149	0.000136	0.002757	0.002718	0.002312	0.00227	0.00375	0.003571
13	47.87013	51	8.762629	9	3.74E-05	3.59E-05	0.000959	0.001011	0.00082	0.00088	0.00139	0.001492
14	14.98715	15	1.997452	2	1.35E-05	1.43E-05	0.00036	0.000359	0.00025	0.00025	0.00044	0.000437
15	5.211481	5	0	0	0	0	0.00013	0.0001	9.28E-05	9.32E-05	0.00018	0.000172
16	0.999583	1	0	0	0	0	5.73E-05	5.74E-05	1.43E-05	1.43E-05	2.87E-05	2.87E-05

Table 3. Comparison of Simulated and Actual Characteristics of Approved Users

Policy Simulations

We conduct several sets of counterfactual simulations to evaluate the effects of eliminating the biases found in the data on the outcome of loan applications across different gender groups. Our counterfactual analyses are done on a second dataset. It covers all the loan records from a one-month experimental period ("full sample" hereafter). During this period, all applicants are approved without screening. As a result, we have true label of all users. This ensures our results are based on the entire user distribution, rather than just the approved users, which have a different distribution from the whole user pool.

Specifically, we calculate the expected profits of the platform based on the predicted approval probability by a number of variants of the estimated structural model. We also examine the gender gap in the approval true positive rate (TPR). We adopt the concept of *equal opportunity*, one of the most popular fairness notions, to measure the extent of bias. *Equal opportunity* requires that qualified individuals, no matter what their sensitive attributes are, have an equal opportunity to receive favorable outcomes. In our loan application setting, this means two gender group should have the same true positive rates. We find that the elimination of either the preference-based bias or the belief-based bias can simultaneously increase the platform's profit and reduce the gender gap in loan approval decisions (TPR).

Does eliminating preference-based bias help with the decision making? How does it affect the platform's payoff?

As noted above, the preference-based bias refers to people's animus towards a particular group. No matter in reality how well this certain group behaves, people with preference-based bias always make judgements and decisions by prejudice. In our setting, we focus on the preference-based bias in gender. In our model, c_{ig} captures the preference-based bias. For applicant i of gender $g \in \{M, F\}$ (M for males and F for females), we normalize c_{iM} for males to be zero. The estimation result for preference-based bias is c_{iF} =-0.2551. Note that in Equation 6, we have a minus sign in front of c_{ig}, therefore c_{iF} =-0.2551 suggests the evaluator has a preference for female borrowers and a prejudice against male borrowers.

	β_F= 0.1133	β_F= 0
c_{iF}=-0.2551	177470.4	180722.5 (+1.83%)
c_{iF}=0	177788.9 (+0.18%)	180954.8 (+1.96%)

Table 4. The expected profits of different decision process.

	β_F= 0.1133	β_F= 0
c_{iF}=-0.2551	13.82%	5.89% (-57.38%)
c_{iF}=0	13.19% (-4.56%)	5.25% (-62.01%)

Table 5. The gender gap in the credit evaluation TPR (female's TPR minus male's TPR).

In an ideal setting, all the evaluators are trained well and prejudice related to gender is completely removed. The elimination of the preference-based bias can be operationalized by setting c_{iF} to be zero. We simulate the evaluators' decisions with $c_{iF} = 0$ but all other parameters unchanged. We then compare the accuracy and the platform's payoff under the original decision-making process and the one with preference-based bias removed on our full sample.

When the evaluators make approval decisions without preference-based bias in gender, the TPR for female users are all lower than their corresponding value from the original decision process. Note that for male users, $c_{iM} = 0$, thus their TPR stays the same under two different decision-making process. In sum, eliminating the preference-based bias can generally decrease the TPR gap between the two gender groups in our setting. But this decrease is relatively small (from 13.82% to 13.19%, Table 5).

We also compare the platform's expected welfare (profit) under the current decision-making process and the one with preference-based bias removed. We observe that by eliminating the preference-based bias (c_{iF}) in the loan approval process, the platform obtains a higher profit (Table 4). The increase in the profit results from better decisions made on female applicants. Specifically, the increase driven by the gain from lowering the approval probability for female borrowers who eventually default on loans, which exceed the loss of lowering the approval probability for nondefault female borrowers.

These findings suggest that although the current decision-making process incorporates gender information to identify high quality users, the evaluators are over-confident about female users and favor female borrowers too much. Therefore, the preference-based bias results in suboptimal decisions.

Does removing gender from the prior belief formation help with the decision making? How does it affect the platform's payoff?

In this subsection, we examine the effects of removing the gender information in the prior belief, i.e. set β_F to be zero. $\beta_g, g \in \{M, F\}$ (M for males and F for females) captures the belief-based bias. Belief-based bias refers to evaluators' subjective beliefs about certain groups. This kind of belief can be updated through observing additional behavioral signals, i.e., repayment behaviors in this paper. In our setting, we normalize male users' β_M to be zero, therefore β_F captures the relative belief-based bias for females compared with males. The estimated value of β_F is 0.1133, which indicates the evaluators have a subjective prior belief in favor of female borrowers. We compare the TPR and the platform's payoff under the original decision process and the one with belief-based bias removed on our full sample.

Under the current decision-making process, we observe a higher TPR for females than males. Since we normalized β_M to be zero, the TPR of males does not change between the two decision processes. Gender information in the prior increases the female TPR. With access to the gender information, the evaluators form a prior belief that favors female borrowers. These additional approved females are generally of good enough credit qualities.

We further investigate the effect of β_F on the expected profit of the platform. With the belief-based bias removed, the platform obtains a larger profit. This suggests that although setting β_F to zero leads to a decrease in the probability of approval for all females, the gain from lowering the approval probability for default female borrowers dominates the loss from lowering the approval probability for nondefault users.

On the borrower side, when the belief-based bias is removed, the gender gap in TPR becomes smaller (Table 5). It decreases from 13.82% to 5.89%. This decrease is larger than the one resulting from eliminating the preference-based bias. When the both biases are ruled out, we can achieve the smallest gender gap in the TPR (5.25%).

Discussion and Conclusions

We have proposed a framework to use structural modeling to distinguish and estimate two types of human bias, i.e., belief-based bias and preference-based bias, based on observational data. In our micro-lending context, the evaluators hold a persistent preference-based bias but learn from three distinct signals (the final overdue days D_{it}, the proportion of installments with positive attitude from the borrower A_{it}, the financial help from family and friends H_{it}), which updates the evaluators' belief-based bias. The model was estimated on real-world data, and our model explains the data well.

The estimation results imply that the evaluators possess a preference-based bias in favor of female applicants and against male ones; they also hold a belief-based bias with a higher prior belief of females' credit qualities. By observing the repayment behaviors, the evaluators can quickly update their belief of the borrowers' credit qualities. And all the three signals play significant roles in the evaluators' learning.

The results from our policy simulations suggest that both the eliminations of the preference-based bias and the belief-based bias can increase the platform's profits. The underlying mechanisms of the two counterfactual settings are the same. Because the loss from lowering the approval probability for nondefault users is smaller than the gain from lowering the approval probability for default users, the platform achieves higher profits. One the borrower side, the eliminations of both types of bias can reduce the gender gap in the credit evaluation true positive rate.

Our paper also has certain limitations that can be addressed in future work. First, the microloan users are generally not stable in their financial condition, which may be one plausible reason why the evaluators heavily rely on the latest repayment behaviors to form a belief of borrowers' credit qualities. In a more stable setting like credit card or mortgage, evaluators may gradually update their beliefs of borrowers' credit qualities. Second, in our policy simulations, we only consider the changes on the evaluator side. In reality, the changes in previous evaluator approval behaviors can also lead to changes in subsequent application behaviors of borrowers. Future work may take both sides into consideration. Third, as a pioneer work on quantifying different types of bias, we do not consider the interaction of gender and other attributes due to model complexity and identification issues. Future work may explore those interaction effects. Despite these limitations, to our best knowledge, this paper is the first to use structural modeling to uncover and distinguish the different types of bias in decision-making processes based on observational data. As machine learning and AI models are increasingly deployed across many decision-making scenarios, it is more and more important to understand the source of biases and propose well-targeted solutions.

Acknowledgement

T. Lu thanks to the support of the National Natural Science Foundation of China (#71872050 and 91546104).

References

Alesina, A. F., Lotti, F. and Mistrulli, P. E. 2013, "Do Women Pay More for Credit? Evidence from Italy," *Journal of the European Economic Association* (11), pp. 45–66.

Arnold, D., Dobbie, W. and Yang, C. S. 2018, "Racial Bias in Bail Decisions," *The Quarterly Journal of Economics* (133:4), pp. 1885–1932.

Bartlett, R., Morse, A., Stanton, R. and Wallace, N. 2019, "Consumer-lending Discrimination in the Fintech Era," *Technical report, National Bureau of Economic Research*.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S. and Roth, A. 2017, "A Convex Framework for Fair Regression," *arXiv preprint* arXiv:1706.02409.

Bertrand, M., Chugh, D. and Mullainathan, S. 2005, "Implicit Discrimination," *American Economic Review* (95:2), pp. 94–98.

Biernat, M. and Manis, M. 1994, "Shifting Standards and Stereotype-based Judgments.," *Journal of personality and social psychology* (66:1), pp. 5.

Blanchflower, D. G., Levine, P. B. and Zimmerman, D. J. 2003, "Discrimination in the Small-business Credit Market," *Review of Economics and Statistics* (85:4), pp. 930–943.

Bohren, J. A., Imas, A. and Rosenberg, M. 2019, "The Dynamics of Discrimination: Theory and Evidence," *American economic review* (109:10), pp. 3395–3436.

Cai, S., Lin, X., Xu, D. and Fu, X. 2016, "Judging Online Peer-to-peer Lending Behavior: A Comparison of First-time and Repeated Borrowing Requests," *Information Management* (53:7), pp. 857–867.

Camerer, C. F. 2019, 24. "Artificial Intelligence and Behavioral Economics", in *The Eco- nomics of Artificial Intelligence*, University of Chicago Press, pp. 587–610.

Chen, D., Li, X. and Lai, F. 2017, "Gender Discrimination In Online Peer-To-Peer Credit Lending: Evidence From A Lending Platform In China," *Electronic Commerce Research* (17:4), pp. 553–583.

Chen, T. and Guestrin, C. 2016, "Xgboost: A Scalable Tree Boosting System, " in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794.

Chen, X., Huang, B. and Ye, D. 2020, "Gender Gap In Peer-To-Peer Lending: Evidence From China," *Journal of Banking and Finance* (11:2), 105633.

Choudhury, P., Starr, E. and Agarwal, R. 2020, "Machine Learning And Human Capital Complementarities: Experimental Evidence On Bias Mitigation," *Strategic Management Journal* (41:8), 1381–1411.

Chouldechova, A. 2017, "Fair Prediction With Disparate Impact: A Study Of Bias In Recidivism Prediction Instruments," *Big data* (52), pp. 153–163.

Corbett-Davies, S. and Goel, S. 2018, "The Measure And Mismeasure Of Fairness: A Critical Review Of Fair Machine Learning," *arXiv preprint* arXiv:1808.00023 .

Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A. 2020, "Biased Programmers? Or Biased Data? A Field Experiment In Operationalizing Ai Ethics," in *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 679–681.

Cui, X. 2019, "Occupational Identity Discrimination In Peer-To-Peer Lending," *Journal of Applied Finance and Banking* (96), pp. 249–278.

Dawes, R. M. 1971, "A Case Study Of Graduate Admissions: Application Of Three Principles Of Human Decision Making," *American psychologist* (26:2), pp. 180.

Dawes, R. M., Faust, D. and Meehl, P. E. 1989, "Clinical Versus Actuarial Judgment," *Science* (243:4899), pp. 1668–1674.

Dobbie, W., Liberman, A., Paravisini, D. and Pathania, V. 2018, "Measuring Bias In Consumer Lending, " *Technical report, National Bureau of Economic Research*.

Drozd, L. A. and Serrano-Padial, R. 2017, "Modeling The Revolving Revolution: The Debt Collection Channel," *American Economic Review* (107:3), pp. 897–930.

Erdem, T., Keane, M. P. and Sun, B. 2008, "A Dynamic Model Of Brand Choice When Price And Advertising Signal Product Quality," *Marketing Science* (27:6), pp. 1111–1125.

Ewens, M. and Townsend, R. R. 2020, "Are Early Stage Investors Biased Against Women?" *Journal of Financial Economics* (135:3), pp. 653–677.

Ferguson, M. F. and Peters, S. R. 1995, "What Constitutes Evidence Of Discrimination In Lending?" *The Journal of Finance* (50:2), pp. 739–748.

Fu, R., Aseri, M., Singh, P. V. and Srinivasan, K. 2019, "'Un'Fair Machine Learning Algorithms," *Available at SSRN* 3408275.

Fu, R., Huang, Y. and Singh, P. V. 2021, "Crowds, Lending, Machine, And Bias," *Information Systems Research* (32:1), pp. 72-92.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T. and Walther, A. 2020, "Predictably Unequal? The Effects Of Machine Learning On Credit Markets," October 1, 2020.

Gneezy, U., List, J. and Price, M. K. 2012, "Toward An Understanding Of Why People Dis- Criminate: Evidence From A Series Of Natural Field Experiments," *Technical report, National Bureau of Economic Research*.

Goldberg, L. R. 1970, "Man Versus Model Of Man: A Rationale, Plus Some Evidence, For A Method Of Improving On Clinical Inferences.," *Psychological bulletin* (73:6), pp. 422.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. and Nelson, C. 2000, "Clinical Versus Mechanical Prediction: A Meta-Analysis.," *Psychological assessment* (12:1), pp. 19.

Hardt, M., Price, E. and Srebro, N. 2016, "Equality Of Opportunity In Supervised Learning," *arXiv preprint* arXiv:1610.02413.

Hu, X., Rudin, C. and Seltzer, M. 2019, "Optimal Sparse Decision Trees," A*dvances in Neural Information Processing Systems NeurIPS* 33.

Huang, Y., Vir Singh, P. and Srinivasan, K. 2014, "Crowdsourcing New Product Ideas Under Consumer Learning," *Management science* (60:9), pp. 2138–2159.

Icard, T. F. 2018, "Bayes, Bounds, And Rational Analysis," *Philosophy of Science* (85:1), pp. 79– 101.

Jadbabaie, A., Molavi, P., Sandroni, A. and Tahbaz-Salehi, A. 2012, "Non-Bayesian Social Learning," *Games and Economic Behavior* (76:1), pp. 210–225.

Kamiran, F., Calders, T. and Pechenizkiy, M. 2010, "Discrimination Aware Decision Tree Learning," in 2010 *IEEE International Conference on Data Mining*, IEEE, pp. 869–874.

KIM, D. 2020, "The Importance Of A Borrower'S Track Record On Repayment Performance: Evidence In P2p Lending Market," *The Journal of Asian Finance, Economics, and Business* 77, pp. 85–93.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Rambachan, A. 2018, "Algorithmic Fairness," in *Aea papers and proceedings* (108), pp. 22–27.

Kleinberg, J., Ludwig, J., Mullainathan, S. and Sunstein, C. R. 2018, "Discrimination In The Age Of Algorithms," *Journal of Legal Analysis* (10), pp. 113–174.

Lambrecht, A. and Tucker, C. 2019, "Algorithmic Bias? An Empirical Study Of Apparent Gender-Based Discrimination In The Display Of Stem Career Ads," *Management Science* (65:7), pp. 2966–2981.

Lin, M. and Viswanathan, S. 2016, "Home Bias In Online Investments: An Empirical Study Of An Online Crowdfunding Market," *Management Science* (62:5), pp. 1393–1414.

Lu, T., Zhang, Y. and Li, B. 2019, "The Value Of Alternative Data In Credit Risk Prediction: Evidence From A Large Field Experiment'.

Lu, Y., Gu, B., Ye, Q. and Sheng, Z. 2012, "Social Influence And Defaults In Peer-To-Peer Lending Networks'.

Lum, K. and Johndrow, J. 2016, "A Statistical Framework For Fair Predictive Algorithms,"*arXiv preprint* arXiv:1610.08077 .

Mateescu, A. 2015, "Peer-To-Peer Lending," *Data and Society Research Institute*.

Parsons, C. A., Sulaeman, J., Yates, M. C. and Hamermesh, D. S. 2011, "Strike Three: Discrimination, Incentives, And Evaluation," *American Economic Review* (101:4), pp. 1410– 35.

Pope, D. G. and Sydnor, J. R. 2011, "What's In A Picture? Evidence Of Discrimination From Prosper. Com," *Journal of Human resources* (46:1), pp. 53–92.

Rudin, C. 2019, "Stop Explaining Black Box Machine Learning Models For High Stakes Decisions And Use Interpretable Models Instead," *Nature Machine Intelligence* (15), pp. 206– 215.

Rudin, C. and Shaposhnik, Y. 2019, "Globally-Consistent Rule-Based Summary-Explanations For Machine Learning Models: Application To Credit-Risk Evaluation," *Available at SSRN* 3395422 .

Rudin, C., Wang, C. and Coker, B. 2020, "Broader Issues Surrounding Model Transparency In Criminal Justice Risk Scoring," *Harvard Data Science Review* 21.

Stevenson, M. T. and Doleac, J. L. 2019, "Algorithmic Risk Assessment In The Hands Of Humans," *Available at SSRN* 3489440 .

Sydnor, J. and Pope, D. 2011, "What'sa Picture? Evidence Of Discriminations Of Loan Fund- Ability In The Prosper. Com Marketplace," *Journal of Human Resources* (46:1), pp. 53–92.

Wang, X., Zhang, D., Zeng, X. and Wu, X. 2013, "A Bayesian Investment Model For Online P2p Lending," *Frontiers in Internet Technologies*, pp. 21–30.

Zhang, S. and Angela, J. Y. 2013, "Forgetful Bayes And Myopic Planning: Human Learning And Decision-Making In A Bandit Setting," A*dvances in Neural Information Processing Systems NeurIPS,* pp. 2607–2615.

Zhang, S., Singh, P. V. and Ghose, A. 2019, "A Structural Analysis Of The Role Of Superstars In Crowdsourcing Contests," *Information Systems Research* (30:1), pp. 15–33.

Zhang, Y., Li, B. and Krishnan, R. 2020, "Learning Individual Behavior Using Sensor Data: The Case Of Global Positioning System Traces And Taxi Drivers," *Information Systems Research* (31:4), pp. 1301-1321.