Northwestern | INSTITUTE FOR POLICY RESEARCH

IPR Working Paper Series

WP-22-53

Smiles in Profiles: Improving Fairness and Efficiency Using Estimates of User Preferences in Online Marketplaces

Susan Athey
Stanford University

Dean Karlan
Northwestern University and IPR

Emil Palikot
Stanford University

Yuan Yuan
Carnegie Mellon University

Version: December 16, 2022

DRAFT

Please do not quote or distribute without permission.

Abstract

Online platforms often face challenges being both fair (i.e., non-discriminatory) and efficient (i.e., maximizing revenue). Using computer vision algorithms and observational data from a microlending marketplace, the researchers find that choices made by borrowers creating online profiles impact both of these objectives. They further support this conclusion with a web-based randomized survey experiment. In the experiment, they create profile images using Generative Adversarial Networks that differ in a specific feature and estimate its impact on lender demand. The authors then counterfactually evaluate alternative platform policies and identify particular approaches to influencing the changeable profile photo features that can ameliorate the fairness-efficiency tension.

A randomized controlled trials registry entry is available at
https://www.socialscienceregistry.org/trials/10030

The authors thank Kiva Microfunds for generously sharing data and discussing the research questions. They thank Herman Donner and Kristine Koutout at Stanford Graduate School of Business and Allison Koenecke at Cornell University for comments and suggestions. This research has been subject to review and approval by Research Compliance Office at Stanford University, protocol number IRB-62442 and registered at AEA RCT registry with the number 0010030.

1 Introduction

Personal profile images play an important role in the success of many online platforms (Ert et al., 2016). At the same time, profile images revealing users' characteristics enable discrimination and can lead to severe inequities in outcomes (Edelman and Luca, 2014). Using data from Kiva, an online micro-lending platform, we examine how profile image features affect fairness and efficiency goals of a platform, and, via simulation exercises, we then examine how platforms can intervene to advance these goals.

To fix ideas, consider sellers on an online marketplace that differ in two dimensions: characteristics that are fixed when they create their online profiles - *type*s and characteristics they choose at this moment - *style*. Suppose that *type* can be high or low and *style* is smiling or not and that buyers prefer sellers of high *type* and sellers with smiling profiles. If *type*s and *style* choices are uncorrelated, we have two distinct sources of inequity: high *type* sellers outperform low *type* sellers and sellers with smiling profiles outperform those without them. When *type* and *style* are positively correlated, the two inequities compound, and, when they are negatively correlated, they mitigate one another.

We analyze *type* and *style* features of online profiles in the context of a non-profit micro-lending platform Kiva. On Kiva, individual lenders make loans to borrowers, selecting from a curated catalog of borrowing campaigns.[1] In principle, Kiva balances two objectives: fairness and efficiency. One way to think about fairness for Kiva is that it involves making access to capital equally available to different *type*s of borrowers. An important component of efficiency is volume, specifically flow of capital from individual lenders (mostly in the United States) to borrowers in low-income countries.

Using observational data on funding outcomes from Kiva, we start by documenting substantial inequities in average daily funds collected by different borrowing campaigns listed on Kiva. Second, we detect features of images using an off-the-shelf machine learning algorithm and select *style* characteristics: features of images that are changeable when users create their profiles. We show that *style* features are predictive of funding outcomes and showcase several specific features that have a large and statistically significant impact on funding outcomes, both unconditionally and after adjusting for other observable profile features. In particular, we show that *smile* is associated with better and *body-shot* with worse funding outcomes. We also document that *style* features are not predictive of the probability of repaying the loan.[2] Finally, we analyze the correlation between *type*s and *style*s. A

[1] Technically, the loan is made to a microcredit institution and earmarked to the specific borrower.

[2] This evidence indicates that the inequity due to *style* features is not justified by different repayment rates. However,

borrower *type* is a collection of characteristics that are fixed at the time the borrower is creating an online profile, such as race, gender, or country of origin. We document patterns of correlation between *type* and *style* characteristics and show that the desirable *style* features are generally more prevalent among borrowers' with *types* associated with better funding outcomes. For example, high-performing borrower types, such as women, are more likely to have profiles with *smile* and less likely to have *body-shot* profile images. This evidence indicates that the distribution of *style* features exacerbates existing inequities in a way that is unfair to the borrowers (not justified by different repayment probabilities).

Estimates of the impact of *style* features on outcomes from observational data rely on the assumption of unconfoundedness, which is not directly testable. In this setting, it requires that for a particular feature such as *smile*, no aspects of the photographs or profile descriptions, correlated with *smile*, other than those that are adjusted for matter for funding outcomes. Even though we use a state-of-the-art feature detection algorithm, it is plausible that we do not capture all information that lenders discern from profile images. To address this issue, we provide evidence from an experiment with recruited subjects on the Prolific.co platform. Subjects choose between profiles of borrowing campaigns featuring fabricated images. The images that we use are generated with Generative Adversarial Networks and can be thought of as pairs of images that are identical except for a feature that we specify. We analyze two features, *smile* and *body-shot*. The estimates that we obtain about the effect of these features on preferences are consistent with our estimates from the observational data from the Kiva platform.

The evidence of the positive correlation between the *types* associated with high funding outcomes and desirable *style* features, indicates that inequities due to intrinsic borrowers' characteristics are exacerbated due to profile *style* choices. However, this also means that outcomes can become fairer if the right policy, which encourages low-*type* borrowers to change their profiles, was implemented. The last part of the paper focuses on comparing the impact of various platform policies on fairness and efficiency.

We consider platform policies that change the conditional distribution of *style* features of borrowers' profiles and that vary probabilities of including borrowers in the lenders' choice set based on borrowers' characteristics. We calibrate a model of lenders' demand for borrowers using estimates from the recruited experiment. We find that policies that alter the distribution of desirable *style* features in the direction that they become less correlated with *type*s improve both fairness and efficiency. Specifically, we show that a policy of *style* curation, which encourages borrowers to have profiles with

style features might be also informative about the developmental impact of the borrowing campaign. We do not adjust the estimates for the expected impact because we do not have a good measure of it; this is a limitation of this work.

smile and avoid *body-shot*s, improves fairness, as measured by the Gini coefficient or the market share of the bottom tercile of borrowing campaigns, and leads to both a higher number of transactions.[3] In contrast, a policy that increases the prominence of campaigns with *smile* and without *body-shot*, for example by ranking them higher on the search page, leads to less fair outcomes, albeit boosts efficiency. This is so because promoting the selected features increases the prominence of high *type* borrowers. Note that if a platform trains a recommendation system based on funding data, and the recommendation system accounts for image features, it is likely that the recommendation system would indeed increase the prominence of profiles with *style* features that are attractive to users, so this policy captures the expected outcome if a platform implements a recommendation system that incorporates image features.

We showcase a specific dimension of *type*-based inequity: the gender gap to the benefit of campaigns with *female* profiles.[4] We show that campaigns with *male* profiles collect 32% fewer funds per day. We corroborate this finding in the recruited experiment where we find that subjects are 31% more likely to choose a *female* profile.[5] The distribution of selected *style* features exacerbates the gap: 77% of borrowers that our algorithm classified as *female*s have profiles with *smile*, as compared to 33% of *male*s; the *body-shot* disparity amounts to 26% to 22%. In the counterfactual simulations, we show that a profile-*style* recommendation policy can substantially narrow the gender gap, while the policy of increasing the prominence of profiles with selected *style* features exacerbates the disparity.

The evidence we present demonstrates how in two-sided markets where users have preferences for features of profile images, the correlation between *type*s and *style* choices can matter for fairness and efficiency. Thus, platforms faced with balancing the two objectives need to account for this correlation before implementing policies based on profile images.

The rest of the paper is organized as follows: Section 2 presents related literature. Section 3 describes how micro-lending platforms operate and provides institutional details about Kiva. Section 4 presents observational data and the evidence obtained from it. Section 5 describes the design of the experiment and its results. Section 6 focuses on counterfactual simulations and Section 7 concludes.

[3] We compare the outcomes under counterfactual policies to a *fair* benchmark in which the distribution of outcomes is unaffected by *style* choices; when the outcomes under a counterfactual policy are closer to the benchmark, we argue that the policy improves fairness.

[4] Throughout the paper we use *male* and *female* to denote the feature detection algorithm's prediction of the gender of the person in the image.

[5] Lenders might prefer campaigns with female profiles for a variety of reasons. For example, there is ample evidence that female entrepreneurs that obtain funding through microfinance generally use the funds effectively (D'Espallier et al., 2011; Aggarwal et al., 2015). Also, lenders might want to compensate for discrimination against women in traditional entrepreneurial finance (Alesina et al., 2013).

2 Literature review

There is rich literature studying the role of images in shaping choices online. In the context of Airbnb, Ert et al. (2016) shows that personal photos increase the sense of personal contact and improve users' perception of the service. Many papers focus on the impact of *type* features. Edelman et al. (2017) and Ge et al. (2016) provide evidence from field experiments documenting that demographic characteristics revealed in images impact the choices of users on hospitality and ride-sharing platforms. In the context of online lending, Theseira (2009); Pope and Sydnor (2011); Younkin and Kuppuswamy (2018) show that loan applications with pictures of black borrowers are less likely to get funded. Jenq et al. (2015) documents that lenders on online peer-to-peer lending favor more attractive and light-skinned borrowers and Ravina (2019) documents an impact of the physical beauty of the borrower. Park et al. (2019) uses an online lab experiment to show that the interaction of borrowers' perceived gender and beauty affects lending decisions. In Kiva's context, Galak et al. (2011) shows experimentally that lenders tend to fund borrowers that are socially similar to them, especially in terms of gender.

Other papers focus separately on the impact of *style* on outcomes. Duarte et al. (2012) shows that borrowers who appear more trustworthy are more likely to have their loans funded.[6] Septianto and Paramita (2021), in a recruited experiment, documents that profiles with happy images receive more donations. Pham and Septianto (2019); Jordan et al. (2019) show that smiling increases the attractiveness of profiles in the charitable giving context. In Kiva's setting, Ai et al. (2016) uses a field experiment document that lenders are more likely to join teams recommended based on location similarity. We contribute to this literature by using Generative Adversarial Networks to provide causal evidence of the impact of selected profile features on outcomes. We introduce a distinction between features of images that are intrinsic to borrowers (*type*) and characteristics that can be modified (*style*) and show that features in both these categories impact outcomes.

We argue that platforms can implement policies that balance fairness and efficiency by exploiting the correlation between desirable *style* features and borrowers *type*s. In this way, the paper relates to the literature analyzing the fairness-efficiency trade-off. There is ample empirical evidence that the implementation of more efficient algorithms can exacerbate inequities (Lepri et al., 2018; Williams et al., 2018; Zhang et al., 2021; Zhang and Yang, 2021). Only a few papers compare various algorithms based on their impact on fairness and efficiency; Rhue and Clark (2020) simulates a marketplace and

[6] Trustworthiness is rated by human-raters based on profile images. Krumhuber et al. (2007) argues that trustworthiness is related to dynamics of facial expressions.

counterfactually adjust algorithmic decision thresholds to highlight the tension between fairness and core business outcomes in an online crowdfunding platform. In the context of criminal sentencing, Kasy and Abebe (2021) presents a theoretical model and calibrated simulations to show how various algorithms impact the race gap. We contribute to this literature by studying a new class of policies based on profile image features. We show that the impact of such policies depends on the correlation between changeable features of images and the fixed characteristics of borrowers depicted in the images. We also demonstrate this point in a counterfactual simulation of various platform policies based on image features.

Our paper showcases that Generative Adversarial Networks can be used to estimate preferences for specific image features. The pipeline that we propose is particularly suitable for audit studies. Audit studies have been commonly used to study biases in how economic agents make decisions (Mullainathan et al., 2012; Kline et al., 2021; Salminen et al., 2022). Our method applies GANs to address the confounding problem by generating images that differ in only the selected dimension; additionally, GAN-generated images are realistic, which allows us to study the effect of the feature in a life-like setting. Fong and Luttmer (2009) varies racial information in images of Hurricane Katrina victims by showing images of black or white hurricane victims. They adjust for other dimensions in which the images differ by reducing image quality and controlling for observed characteristics. Ash et al. (2022) uses the interaction between the text of a news article and the associated image to measure the extent of racial stereotypes. Flores-Macías and Zarkin (2022), in a conjoint experiment, studies the effect of military uniform, gender, and skin color on the perception of the effectiveness of law enforcement. The images used in the experiment are modified using Photoshop. GANs improve upon these methods by varying only the selected feature and producing high-quality realistic images; they can also be scaled to a larger set of images at low cost. Another paper that uses GANs for a related purpose is Ludwig and Mullainathan (2022). The key difference between the two papers is the specific objective for which the GANs are used. Ludwig and Mullainathan (2022) uses GANs to morph images in a direction that shifts the choices of decision-makers and then asks recruited subjects to name the changed feature. In our paper, we start by identifying features, prioritize those for study that appear to have a causal effect in observational data, and finally modify images with respect to a selected feature and use GANs to estimate the impact of the feature on the choices of recruited subjects.

3 Empirical context

Microcredit, sometimes called microlending, is amorphously defined but typically refers to small uncollateralized loans to low-income households at terms more favorable than otherwise available, often on the premise of supporting micro-enterprise development.[7]

Over the past two decades, online microcredit platforms have broadened participation opportunities by creating access for individual lenders. Sun et al. (2019) argues that online microcredit platforms, such as Kiva, help establish personal connections between lenders and borrowers and encourage participation by simplifying the discovery and lending processes.

While the microcredit platforms have enabled the participation of new lenders, there have also been increased concerns over issues of fairness. Past research using field and lab experiments has shown that online peer-to-peer platforms yield considerable inequities across race, gender, and physical attributes of borrowers.[8]

A number of microcredit platforms are non-profit organizations, including Kiva. In their context, the tension between fairness and efficiency is particularly nuanced. On the one hand, improving efficiency increases the number of borrowers who are otherwise financially excluded from the source of credit; but on the other hand, it changes the dynamics of fund access and might create inequality among different types of borrowers.

Kiva. Serving borrowers in more than 80 countries, Kiva is one of the most prominent online non-profit peer-to-peer microcredit platforms. Since its founding in 2005, Kiva has issued over 1.6 million loans funded by over 2 million lenders totaling 1.7 billion U.S. Dollars. Kiva creates an online marketplace where borrowers have dedicated profile pages with pictures that prospective lenders can browse to select the ones they want to invest in. Kiva collaborates with local microcredit agencies in vetting, curating, and promoting borrowers.

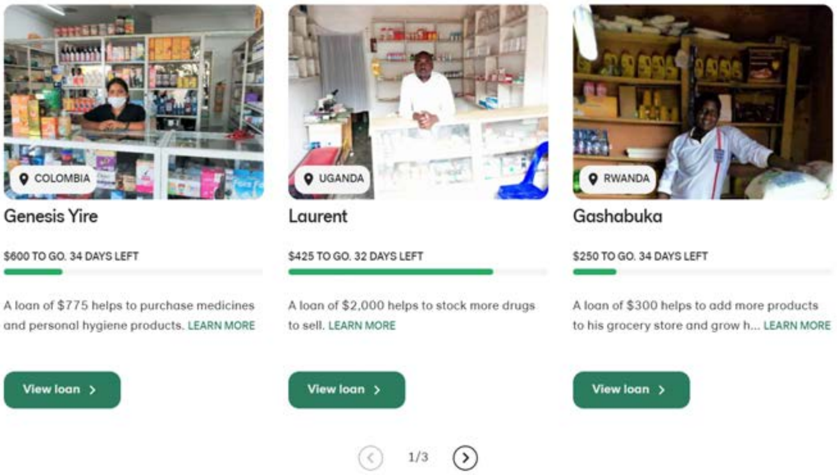
A potential lender makes the first contact with a campaign through the category page. Figure 1 shows an example. Lenders can obtain more information by clicking on "View loan," where they learn more about the loan objective and geographical location.

Images play a prominent role in lenders' discovery of borrowers and they help borrowers in

[7] See Karlan and Morduch (2009) for an academic overview of microcredit, and see Banerjee et al. (2015) for a summary of six randomized controlled trials of microcredit.

[8] Fong and Luttmer (2009) documents that lenders prefer borrowers of the same race; Landry et al. (2006); Park et al. (2019) show the role of gender and perceived beauty.

Figure 1: Kiva category page.



Genesis Yire

$600 TO GO. 34 DAYS LEFT

A loan of $775 helps to purchase medicines and personal hygiene products. LEARN MORE

Laurent

$425 TO GO. 32 DAYS LEFT

A loan of $2,000 helps to stock more drugs to sell. LEARN MORE

Gashabuka

$250 TO GO. 34 DAYS LEFT

A loan of $300 helps to add more products to his grocery store and grow h... LEARN MORE

Note: Screenshot from kiva.org collected on 3/3/2022.

telling their stories. Lenders use it as a key factor when deciding whether or not to invest (Park et al. (2019)). Borrowers take and provide their own images. As a result, profile images differ in their content. Some show mostly the borrower while others focus on their business; facial expressions of borrowers, e.g., serious or smiling, and technical aspects such as quality of lighting or resolution tend to vary too. In order to help the borrowers with this important part of their application, Kiva.org provides some suggestions about how profile photos should look, e.g., they recommend that photos should be in high resolution and horizontal orientation and include the business owner and the business in the background.[9]

4 Analysis of observational data

Our framework for studying fairness and efficiency requires that we establish that lenders have preferences for specific profile *style* features and that they are differentially distributed across borrowers *type*s. To do that we use data on the historical performance of borrowing campaigns on Kiva.

4.1 Kiva data

We construct *Kiva data* by combining three datasets: a publicly available dataset with loan characteristics and lending outcomes, data on features of images associated with the borrowing campaigns obtained with the methodology described in Appendix A, and a dataset on repayments that has been generously shared by Kiva.[10]

[9] See https://www.kivaushub.org/profile-photo

[10] See here: https://www.kiva.org/build/data-snapshots for the publicly available dataset.

The publicly available data on characteristics and outcomes of borrowing campaigns spans April 2006 until May 2020 and contains over half a million observations. Data describe key features of each borrowing campaign such as sector, name of activity, country, funding goals, or currency. We have several metrics of funding outcomes: the amount of money collected per day, the number of days it took to raise the capital (campaigns generally stay active until they collect all funds), and the number of lenders that loaned money to the borrower. We primarily focus on money collected per day as an outcome metric because we are interested in analyzing how lenders allocate their capital between borrowers to estimate their preferences for specific features of borrowing campaigns. We characterize the competitive landscape by exploiting the fact that our data contains all borrowers available in the covered period. For each borrower, we compute the number of borrowers from the same country and sector listed concurrently. We also include the share of borrowers of the same *race* and *gender*.[11] Finally, to flexibly capture time trends, we introduce interactions between the month in which a campaign was posted and the sector and interaction between the country and the month.

Images associated with borrowing campaigns are also publicly available. We use the feature detection algorithm: Convolutional Neural Network (CNN), described in the Appendix A, to generate features of profile images and enrich the funding outcomes dataset. The algorithm that we use takes as input an image and returns a vector of probabilities associated with a pre-defined list of features. From the CNN, we obtain around 140 features of images: various objects in the image, technical aspects of the photo (*blurry*, *flash*), facial expressions of a person, and other individual characteristics like *race* or *age*. However, not all of these features are useful in our context. First, many of the features do not or very rarely appear in our dataset. To reduce data size, we remove features that take the value of zero for more than 99.9% of images. By doing so, we mostly drop features describing specific objects in the image (e.g., *cup*). Second, we drop several features that are highly correlated (e.g., *frowning* and *smiling*) since such features mostly duplicate information.[12] In the end, we focus on 55 features of images. The full list is available in AppendixA.

The feature detection algorithm returns an estimate of the probability that the specific feature is present in the image; depending on the application, we either use the continuous value or a binary indicator taking the value of one when the probability exceeds 50% and zero otherwise. We use *italics* when referring to demographic features predicted using CNN.

Out of the 55 image characteristics, some cannot be changed when borrowers create their profiles

[11] *Race* and *gender* are predictions based on campaigns' profile images.

[12] When several features have the Pearson correlation coefficient above 0.75, we select one of them.

(e.g., demographics).[13] We categorize such features as *type*. Formally, we treat *types* as a collection of features both from a profile image as well as from the description of the borrowing campaign (e.g., country). In contrast, features that borrowers can change when creating their profiles we label as *style*. We categorize the following image characteristics as *style*: *No Eyewear*, *Sunglasses*, *Smile*, *Blurry*, *Eyes Open*, *Mouth Wide Open*, *Blurry*, *Harsh Lighting*, *Flash*, *Soft Lighting*, *Outdoor*, *Partially Visible Forehead*, *Color Photo*, *Posed Photo*, *Flushed Face*, *top* (person's face in the top part of the image), *right* (person's face in the right part of the image), *bottle* (there is a bottle in the image), *chair* (there is a chair in the image), *person* (there is another person in the image), and *body-shot* (body of the borrower occupies a substantial part of the image).[14]

Finally, the data on defaults span from 2006 until 2016 and contains approximately 420 thousand borrowing campaigns. The unit of observation in this dataset is a loan of an individual lender; note, generally, borrowing campaigns have multiple lenders. We have information on whether each loan has been repaid or not.[15] We aggregate the dataset to the borrowing campaign level and consider the loan to be repaid if the borrower paid back the money to all lenders. Thus, the variable of interest is whether borrowers defaulted or paid all their loans.

We merge the three datasets and as a result, the final dataset captures the period from 2006 to 2016. Table 1 presents summary statistics of the main variables. The full list is available in Appendix 6.

4.2 Inequities in funding outcomes

While loans on Kiva stay active for a long time so that the vast majority of them eventually get funded, there is a substantial variation in how long it takes to reach campaigns' funding goals or how much money is collected per day. In Figure 2, we show a histogram of the number of days it takes to collect the entire amount (*days to raise*) and a Lorenz curve documenting inequity in this outcome. If every borrowing campaign would take the same number of days to get funded, blue (actual distribution) and gray (perfect equality) curves would overlap.

From the left panel of Figure 2 we can observe that there is substantial dispersion in how long it

[13] Of course, it may be possible for borrowers to shift perceptions to create ambiguity about their type features; in this paper, we abstract from such manipulation, but caution that if platform participants learn that there is a benefit to doing so, they may indeed engage in such behavior.

[14] Body-shot: a type of camera shot in which a character's body is the primary content of the image and reaches from the top of the frame to the bottom of the frame.

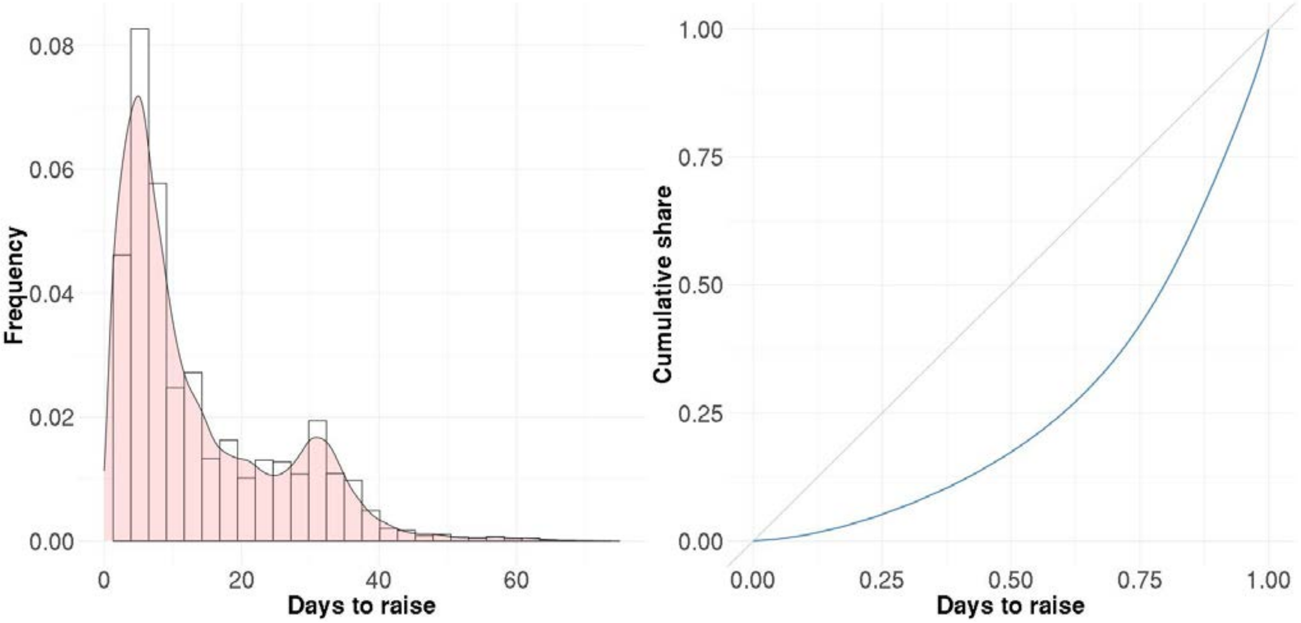
[15] We also have information on who defaulted on the loan: (i) defaults by the borrower 75% of cases, (ii) defaults by the micro-finance partner 23% of cases, and both 2% of cases. Each of these categories has the same impact on the lender, the loan is not repaid. Thus, in the main analysis, we do not distinguish between the reasons for the default. See Appendix Efor further discussion.

Table 1: Summary statistics of the main variables.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
cash per day	420,765	104.587	136.378	1	25	116.7	621
days to raise	420,765	13.175	10.947	1	5	20	38
default	420,765	0.050	0.218	0	0	0	1
loan amount	420,765	800.107	993.370	25	275	950	50,000
no. competitors	420,765	0.091	0.173	0.003	0.006	0.075	1.000
share same race and gender	420,765	0.665	0.294	0	0.4	1	1
male	420,765	0.198	0.398	0	0	0	1
smile	420,765	0.498	0.177	0	0	1	1
body-shot	420,765	0.406	0.491	0	0	1	1

Note: Summary statistics of selected variables. Cash per day and days to raise winsorized at top 97th percentile. Cash per day and Loan Amount in USD dollars; male and smile take the value of 1 when CNN predicted probability is above 0.5 and zero otherwise; body-shot takes the value of one when the body of the person takes more than 33% of the area of the image. No. competitors the number of borrowing campaigns from the same sector and country posted concurrently, the value is standardized by the maximum.

Figure 2: *Days to raise*: histogram and Lorenz curve.

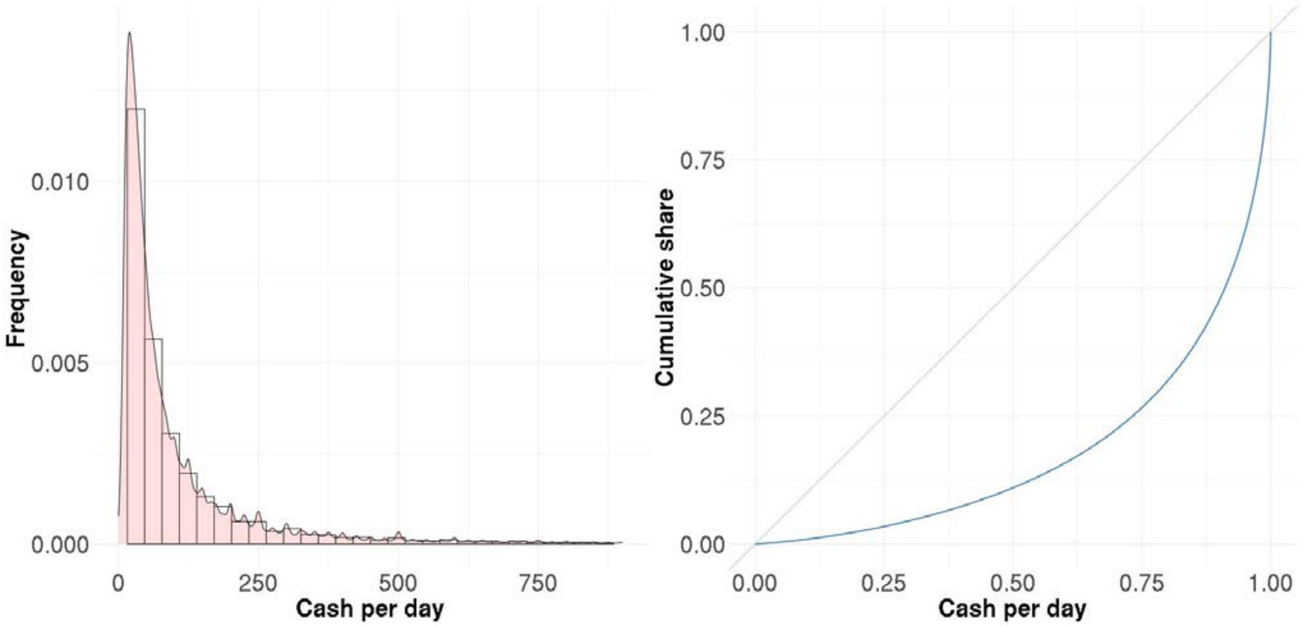


Note: Left panel - histogram of days to raise capped at 75 USD. In pink a fitted density curve. Right panel - Lorenz curve of days to raise.

takes to collect the entire amount of the loan. The mean outcome is 14.5 days, but many campaigns fill in almost instantaneously while others take over a month to reach their funding goals.

An important driver of how long it takes to collect the entire amount is the size of the loan. Thus, a useful measure of how quickly borrowers raise funds is the amount of money raised per day. In Figure 3, we show a histogram of funds in dollars collected per day (*cash per day*) and an associated Lorenz curve. There is a substantial variation in *cash per day*. The mean amounts to 118 USD, but many campaigns raise just a few dollars per day. Focusing on the Lorenz curve (right panel) we observe even higher inequities than in Figure 2.

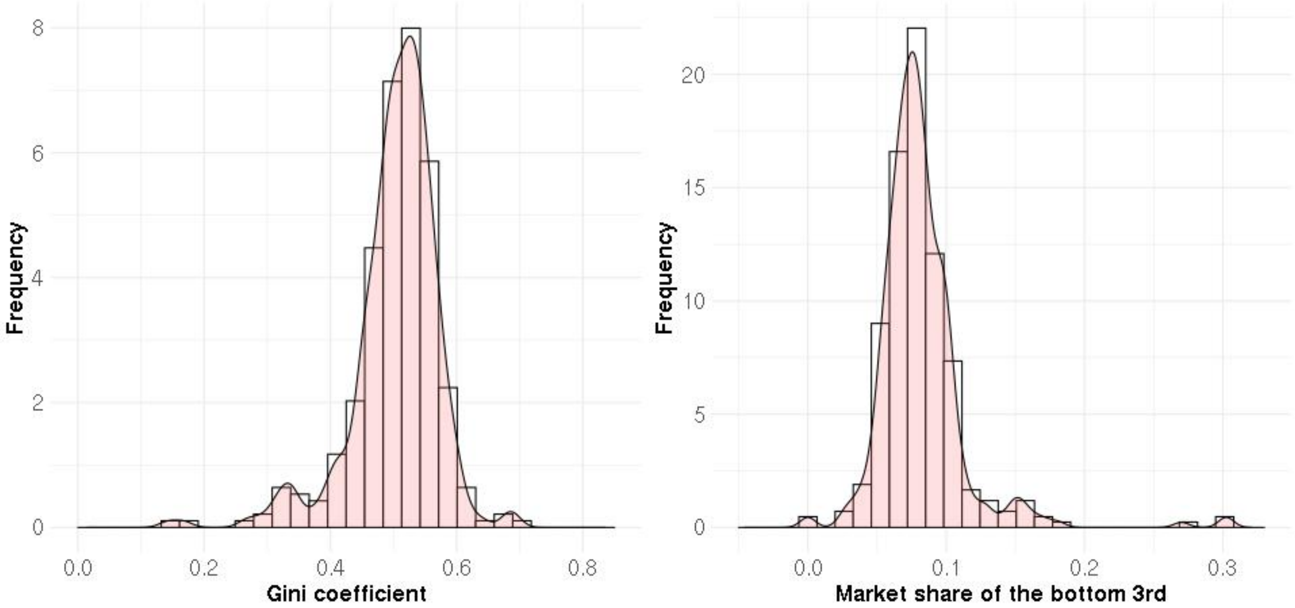
Figure 3: *Cash per day*: histogram and Lorenz curve.



Note: Left panel - histogram of cash per day capped at 1250 USD. In pink a fitted density curve. Right panel - Lorenz curve of cash per day.On average there are 450 borrowing campaigns available in a given week, which together raise on average over USD 400,000 in loans.

The evidence presented in Figures 2 and 3 is based on data collected over ten years. Much of the variation can be due to time trends such as differences in the number of available lenders or borrowers. In Figure 4, we group campaigns into weekly intervals such that campaigns that were available online during the same week are in the same group. Thus, a group of borrowing campaigns approximates a choice set available to lenders that were active in that week. We use two measures of inequity the Gini coefficient and the sum of market shares of the 33% of borrowers with the lowest amount of money collected per day.[16] The Gini coefficient of 0 expresses perfect equality, where all values are the same. The market share of the bottom third amounting to 33% would indicate that the outcomes as equally distributed across tertiles. We can observe that both metrics document that outcomes are far from being equally distributed.

Figure 4: *Cash per day* distribution within weeks: Gini coefficient and share of the bottom tertile.



Note: Statistics in both panels computed on a weekly basis. Left panel - Gini coefficients of weekly distributions of cash collected per day. Right panel - weekly sums of cash collected per day by the 33% lowest performing borrowers.

4.3 Profile images and outcomes.

Funding outcomes. Evidence presented in Figures 2, 3, and 4 indicates that there are substantial inequities in funding outcomes across borrowing campaigns. The inequities can be due to differences in borrowers' *type*s and their *style*. *Style* features are central to this analysis because a platform can design interventions modifying them. In Table 2, we show that part of the variation in outcome can be explained by image *style* features.

We train three models to predict *cash per day*, first, a full model which includes all variables in *Kiva data*, second, a restricted model that contains only *style* features, and finally, a benchmark mean

[16] The Gini coefficient defined as

$$Gini = \frac{\sum_{j=1}^{n} \sum_{j'=1}^{n} |x_j - x_{j'}|}{2n\bar{x}}$$

where x_j is the outcome for borrower j and $x_{j'}$ for borrower j', n is the number of borrowers available in that week and $\bar{x}$ the average cash collected per day.

model. We use a gradient boosted machine (Friedman (2001)) for the predictive model.[17] We split the dataset 70:30 into train and test. Table 2 reports predictive performance in the test set measured using mean squared error. We find that including *style* features improves the predictive performance of the model, as compared to a mean model. Additionally, a full model based on all covariates in *Kiva data* performs better than the model with only *style* features.

Table 2: Image features as predictors of *cash per day*.

specification	MSE	SE
Mean	22367	252
Style features	19373	224
Full model	10996	138

Note: Test set performance of a gradient boosted machine (GBM) trained using all available covariates (full model) models with only image style *features and a mean model. Models trained on 70% of data and tested on 30%. In the second column, the mean squared errors. Third column standard error of MSE.*

Specific style features. Results presented in Table 2 show that both *style* and other image features altogether are predictive of funding outcomes. However, to construct platform policies around *style* features it is important to select individual impactful features. In other words, we want to ask a question: "what would happen if a profile was presented with a change in one characteristic and remained unchanged otherwise." Therefore, the estimate of our interest is the average treatment effect (ATE) of a specific image feature.

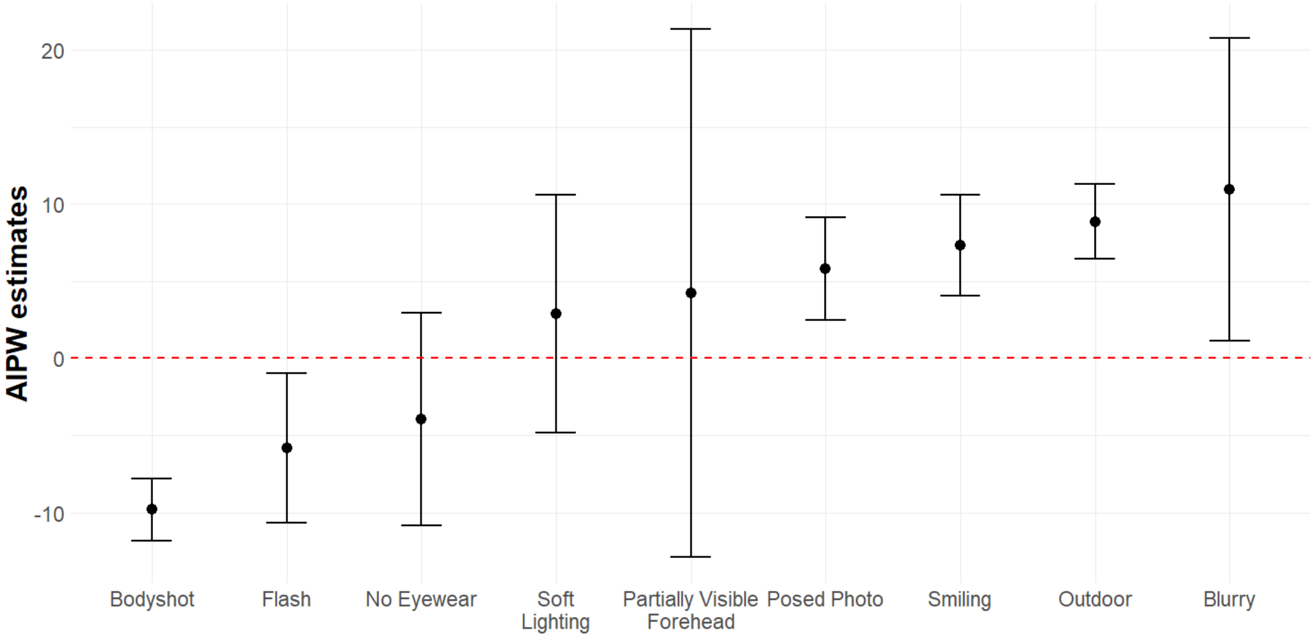
To estimate ATEs we use Augmented Inverse Propensity Weighing (AIPW) estimator (Robins et al., 1994; Glynn and Quinn, 2010; Wager and Athey, 2018). AIPW is a doubly-robust method; thus, it adjusts for covariates in the outcome model and the propensity score. We use the *grf* implementation of the AIPW estimator (Athey et al. (2019)).

The feature detection algorithm detects a very rich set of characteristics, however, possibly there are still some variables (both image feature and other) that we are missing, which means that there still might be other factors that influence lenders' decisions. Thus, the ATE estimates should be interpreted as comparing profiles that are similar in all observed dimensions other than the studied one. We return to this issue in Section 5, where we present experimental evidence corroborating the impact of selected features on outcomes.

Figure 5 shows estimates of average treatment effects on *cash per day* for selected features. Results

[17] Unless stated otherwise we use the gradient boosted machine for all predictive tasks. We selected the model based on a comparison of test set performance with other popular predictive models. See Appendix D for a detailed discussion.

Figure 5: Estimates of the average treatment effect of selected style features



Note: Estimates of the average treatment effect of selected features on cash collected per day. x-axis selected features, y-axis ATE estimates. 99.9% confidence interval. Propensity and outcome model estimates using Regression Forest. We transform the treatment variable to a binary variable that takes the value of one when the predicted probability of the feature is above 0.5 and zero otherwise. See Appendix F for results using GBM and diagnostics.

for all features, diagnostics, and other estimators are available in Appendix F. We find that several features have negative ATE e.g., *flash*, *body-shot*, while others like *posed photo* or *smile* have positive and statistically significant effects.

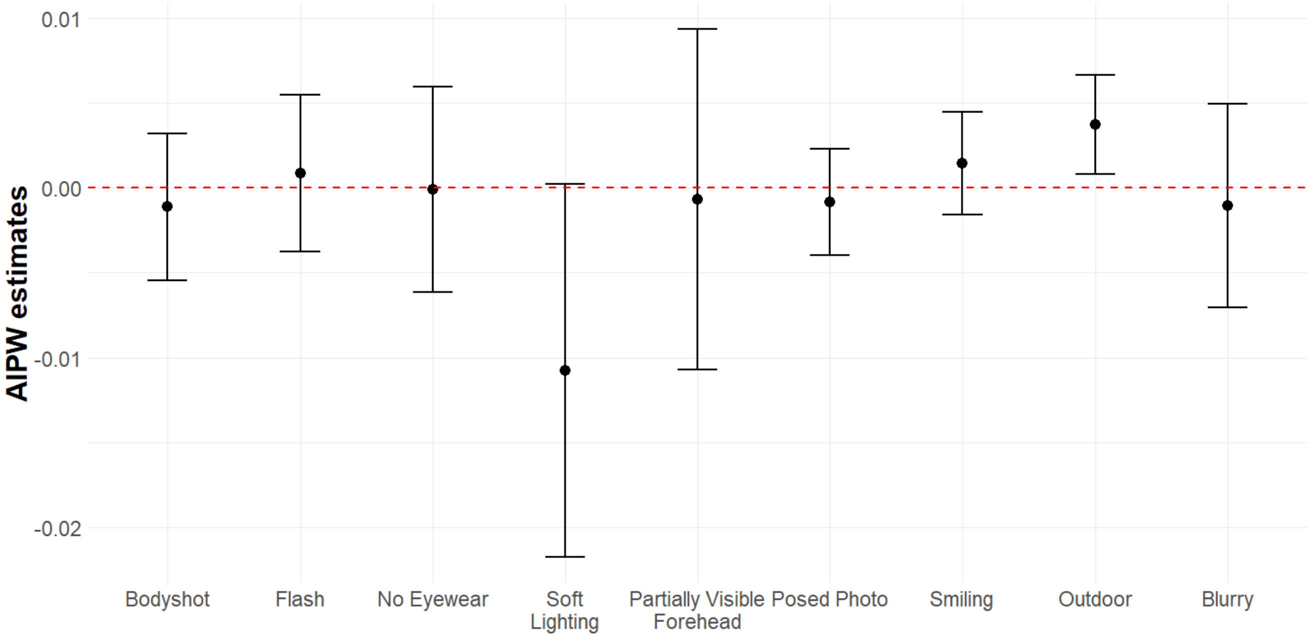
Loan repayment. Lenders might use *style* features as signals of the probability of repayment; we show that *style* features are not predictive of the probability to repay the loan. In Table 3, we compare the predictive performance of default models with and without image features. We see that the inclusion of *style* features does not improve the predictive performance of the model as compared to a mean model.

Table 3: Image features as predictors of default probability.

specification	MSE	SE
Mean model	0.065	0.001
Style features	0.064	0.001
Full model	0.059	0.001

Note: Test set performance of a gradient boosted machine (GBM) trained using all available covariates (full model) and simplified models image style features (Style features) and a model with only an intercept (Mean model). Models trained on 70% of data and tested on 30%. In the second column, the mean squared errors. Third column standard error of MSE.

Figure 6: Estimates of the average treatment effect of selected style features



Note: Estimates of the average treatment effect of selected features on probability to default. x-axis selected features, y-axis ATE estimates. 99.9% confidence interval. We transform the treatment variable to a binary variable that takes the value of one when the predicted probability of the feature is above 0.5 and zeroes otherwise.

We also estimate the average treatment effects of individual features on the probability of default. Figure 6 shows the results. The only feature associated with a statistically significant impact on repayment probability is *outdoor*. We take this results as evidence that *style* features are not useful signals of repayment probability.

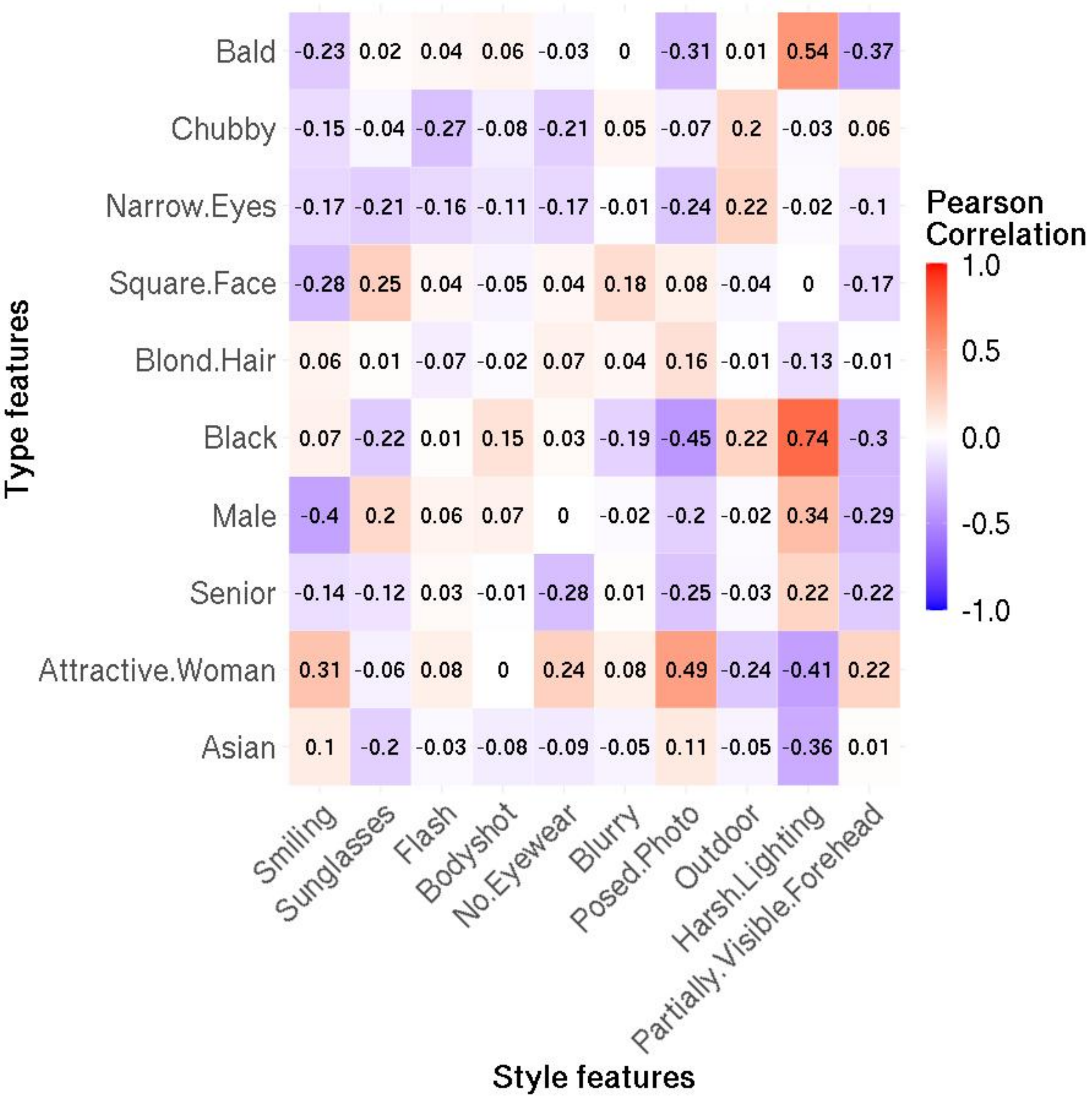
4.4 Correlation of *type*s and *style* features

Style features can aggravate or mitigate inequities in outcomes due to differences in *type*s. When borrowers with *type* feature associated with high outcomes build profiles that attract lenders, the disparities due to *type*s will increase further; in contrast, if borrowers with less desirable *type* features choose attractive profile *style*s, outcomes will be more equitable. In this section, we document the correlation between *type*s and *style*s.

Figure 7 shows a correlation between selected *type* and *style* features. Some of the features are highly correlated; for example, *smiling* is less prevalent amongst *male* and *bald* and more common for *attractive woman*.

To argue that the choices of *style* features exacerbate inequities due to *type*s, we need to show that the distribution of *type*s across borrowing campaigns results in inequity of outcomes and that the

Figure 7: Correlation between selected *type* and *style* features.



	Smiling	Sunglasses	Flash	Bodyshot	No.Eyewear	Blurry	Posed.Photo	Outdoor	Harsh.Lighting	Partially.Visible.Forehead
Bald	-0.23	0.02	0.04	0.06	-0.03	0	-0.31	0.01	0.54	-0.37
Chubby	-0.15	-0.04	-0.27	-0.08	-0.21	0.05	-0.07	0.2	-0.03	0.06
Narrow.Eyes	-0.17	-0.21	-0.16	-0.11	-0.17	-0.01	-0.24	0.22	-0.02	-0.1
Square.Face	-0.28	0.25	0.04	-0.05	0.04	0.18	0.08	-0.04	0	-0.17
Blond.Hair	0.06	0.01	-0.07	-0.02	0.07	0.04	0.16	-0.01	-0.13	-0.01
Black	0.07	-0.22	0.01	0.15	0.03	-0.19	-0.45	0.22	0.74	-0.3
Male	-0.4	0.2	0.06	0.07	0	-0.02	-0.2	-0.02	0.34	-0.29
Senior	-0.14	-0.12	0.03	-0.01	-0.28	0.01	-0.25	-0.03	0.22	-0.22
Attractive.Woman	0.31	-0.06	0.08	0	0.24	0.08	0.49	-0.24	-0.41	0.22
Asian	0.1	-0.2	-0.03	-0.08	-0.09	-0.05	0.11	-0.05	-0.36	0.01

Note: Pearson correlation coefficient between selected style *features in columns and* type *features in rows.*

desirable *style* features are more prevalent amongst borrowing campaigns whose *type*s lead to better funding outcomes.

To do that we carry out a *Gelbach Decomposition* (Gelbach, 2016) of selected *type* variables. This method allows measuring to what extent adjusting for a group of variables changes the coefficient of a selected variable and informs us what the coefficient would look like if the means of adjusting variables were the same across the levels of the evaluated variables.

Table 4 presents the results for a selected *type* variables. The first column shows the name of the variable. The second column shows the coefficient associated with the selected variable from a linear regression of the variable on *cash per day*. In the third column, we see the coefficient adjusted for all variables in *Kiva data*. In the final column, we have the adjustment due to *style* features. For example, if we partial out differences in distribution of *style* features profiles with *bald type* make USD 10.21 less than those of not *bald type*. Thus, deferentially distributed *style* features aggravate disparity between *bald* and not *bald*. Finally, as evidenced in Table 3, the additional inequity is not explainable by differences in repayment probability, and in this sense is unfair to the borrower.

Table 4: Impact of style features on coefficients associated with *type*s

feature	Coefficient base	Std. error base	Coefficient full	Std. error full	Delta style
Bald	-71.57	4.86	-22.14	7.31	-10.21
Chubby	16.92	2.02	-7.65	4.17	2.08
Narrow Eyes	-9.14	1.87	3.77	3.66	2.01
Square Face	-87.70	8.06	-27.64	10.46	-52.59
Black	-12.06	1.37	-1.61	3.51	4.61
Senior	-57.18	4.70	-6.02	5.96	-6.02
Attractive Woman	75.32	2.43	10.72	4.13	9.48
Asian	12.39	1.52	0.72	2.74	2.30

Note: Gelbach decomposition of selected type *features (Gelbach, 2016).* Coefficient base *refers to coefficient of a univariate model with the selected* type *feature;* coefficient full *coefficient from a model adjusting for all covariates in* Kiva data; delta style *impact of style features on the disparity between* types. *R implementation by Stigler (2018)*

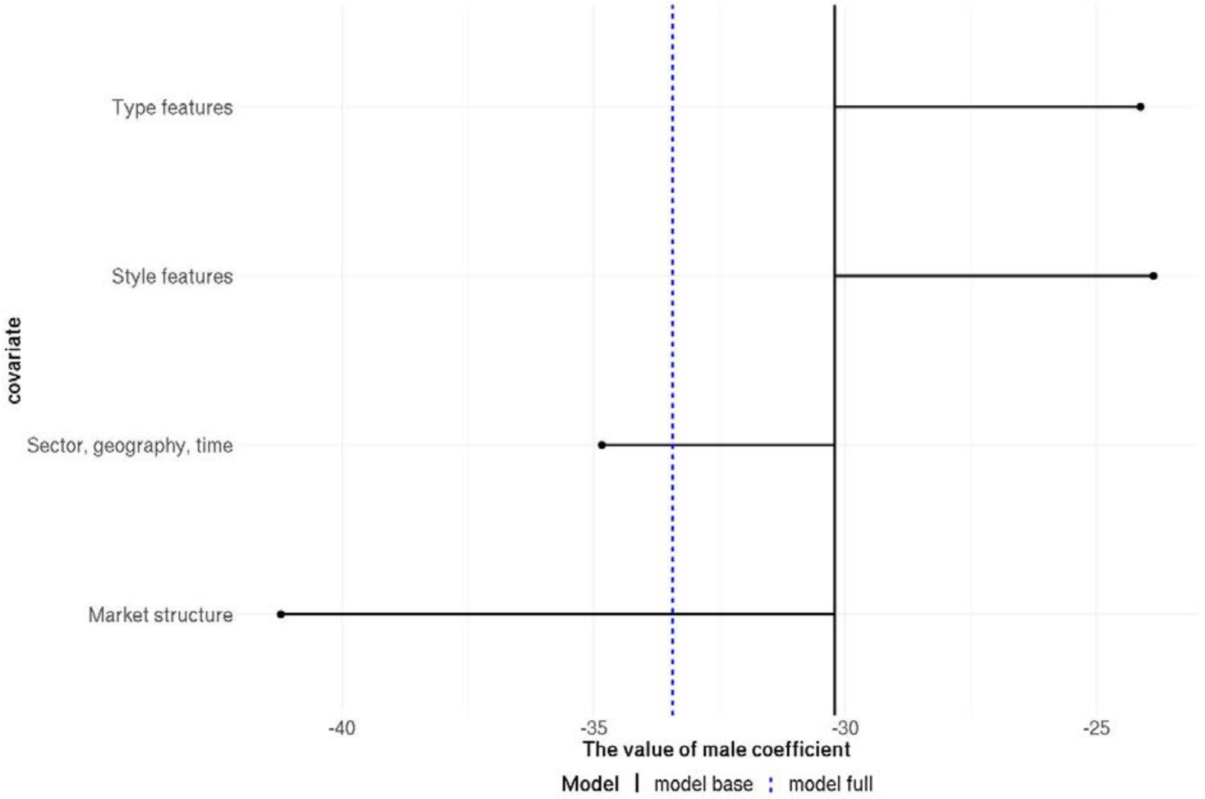
Results presented in Table 4 indicate that *style* choices aggravate disparities due to *bald, square face, senior, attractive woman, and Asian types* and mitigate for *chubby, narrow eyes, and black types*.

4.5 *Gender* gap

A specific *type* feature that matters in our context is *gender*.[18] We find that campaigns classified as *male* raise on average USD 36 less per day and take 5.8 more days to achieve funding objectives (differences

[18] We use an algorithmic prediction of *male*. Thus, the variable *male* indicates that the feature detection algorithm assigns the probability of at least 0.5 that the person in the image is a male.

Figure 8: Gelbach decomposition of *male* coefficient



Note: Solid line estimate of the coefficient associated with male *from a univariate linear regression; dashed line the coefficient adjusted for all variables in Kiva data; OLS estimator. Horizontal lines represent contributions of variables group to the coefficient associated with* male; type *features include all other* type *features from the image;* sector, geography, time *includes* sector, country, week *fixed effects,* loan amount, *and* repayment details; *and* market structure *includes interactions of* month *and* country, month *and* sector, number of lenders in the week, number of competing campaigns, *and* share of campaigns of the same race and gender. *The R implementation of Gelbach (2016) by Stigler (2018).*

in means).[19]

As evidenced by the results presented in Table 4, the differences between *types* can be due to the non-equal distribution of other characteristics, e.g., *style*. To decompose the unadjusted difference we again carry out a *Gelbach Decomposition* (Gelbach, 2016); the proposed method compares a baseline model with just *male* indicator variable with a full model that includes all the variables in *Kiva data* and decomposes the contribution of all the added variables to the changes in the coefficient of interest. Figure 8 presents the results.

Figure 8 depicts the differences in the coefficient associated with *male* between a univariate linear

[19] In the context of microfinance, the gender gap might be driven by users that aim to correct for discrimination against women in traditional finance. There is a rich literature documenting discrimination against women in traditional entrepreneurial lending. Alesina et al. (2013) shows that women entrepreneurs pay higher rates for access to credit and Brock and De Haas (2021) use a randomized experiment to show that loan officers grant loans to women under less favorable conditions than to men. The phenomenon of over-correcting for discrimination is well documented in experimental psychology (Mendes and Koslov (2013), Nosek et al. (2007)).

regression (solid line) and a full model which includes all variables from *Kiva data* (dashed line). The length of the horizontal arms going from the base model indicates the contribution of each variable group to the *male* coefficient in the full model. Thus the length of the horizontal line is the partial effect of the unequal distribution of features within the group. We can observe that changing distributions of *style* features would decrease the gender gap; additionally, we find that *male* campaigns have also a non-desirable distribution of other *type* features, but the distribution of *sector, geography, time* and *market structure* decreases the *gender* gap.

Additionally, we also look at the prevalence of the desirable and non-desirable features. For example, the frequencies of *body-shot* and *smile* differ substantially between genders: 77% of *female* borrowers *smile* in the image as compared to 33% of *male*. 26% of *male* use *body-shot* as compared to 22% of *female* (both are statistically significant).

5 Recruited experiment

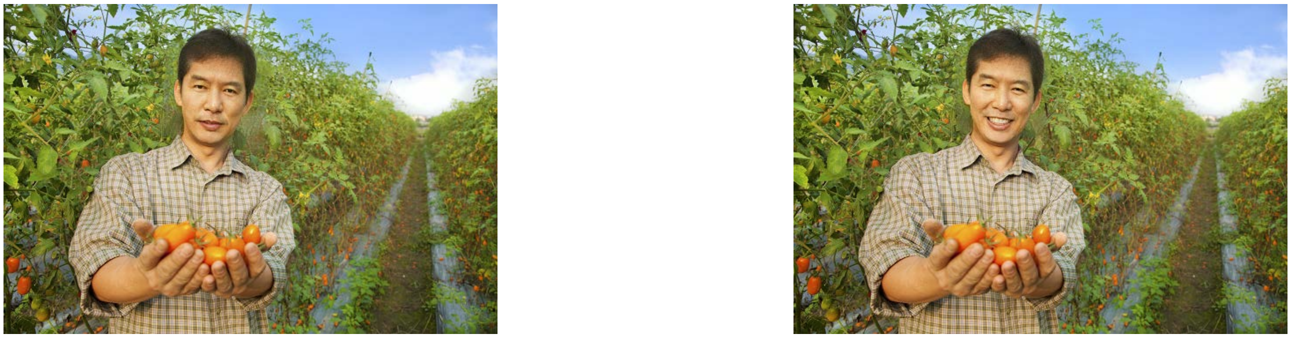
The key components of our conceptual framework are the distribution of *style* features across high and low-performing borrowers and the effects of these features on outcomes. The causal interpretation of the estimates of treatment effects from Section 4 rests on the assumption of unconfoundedness, which we cannot verify. This section provides experimental evidence of the impact of two selected style features: *smile* and *body-shot* on outcomes. We selected these features because of high and statistically significant estimates of their impact on *cash per day*, high correlation with borrowers' types, and differences in their prevalence amongst *male* and *female*.

5.1 Experiment design

In the experiment, subjects are presented with a series of pairs of borrowers and asked to select one out of each pair. Pairs of profiles feature fabricated images with exogenous variation in *smile*, *body-shot*, and *male*. Thus, images differ in the three dimensions and the objective of the experiment is to estimate the treatment effects of *smile*, *body-shot*, and *male*. The design of the experiment builds on the literature on conjoint analysis (Hainmueller et al. (2014)) with the novelty that variation in features of interest is encoded in images.

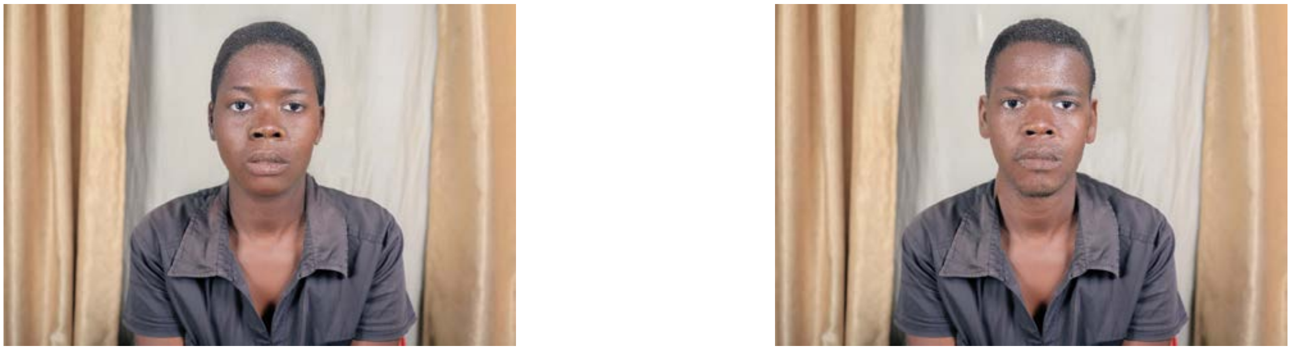
Images. To generate images that differ in the selected features we use Generative Adversarial Networks (GANs). GANs designed by Goodfellow et al. (2014) is an approach to generative modeling using deep-learning methods. The key objective of GANs is to generate fabricated data that are sim-

Figure 9: Variation in *smile*



Note: Two versions of an image with variation in smile. Both images were generated using GANs.

Figure 10: Variation in *male*

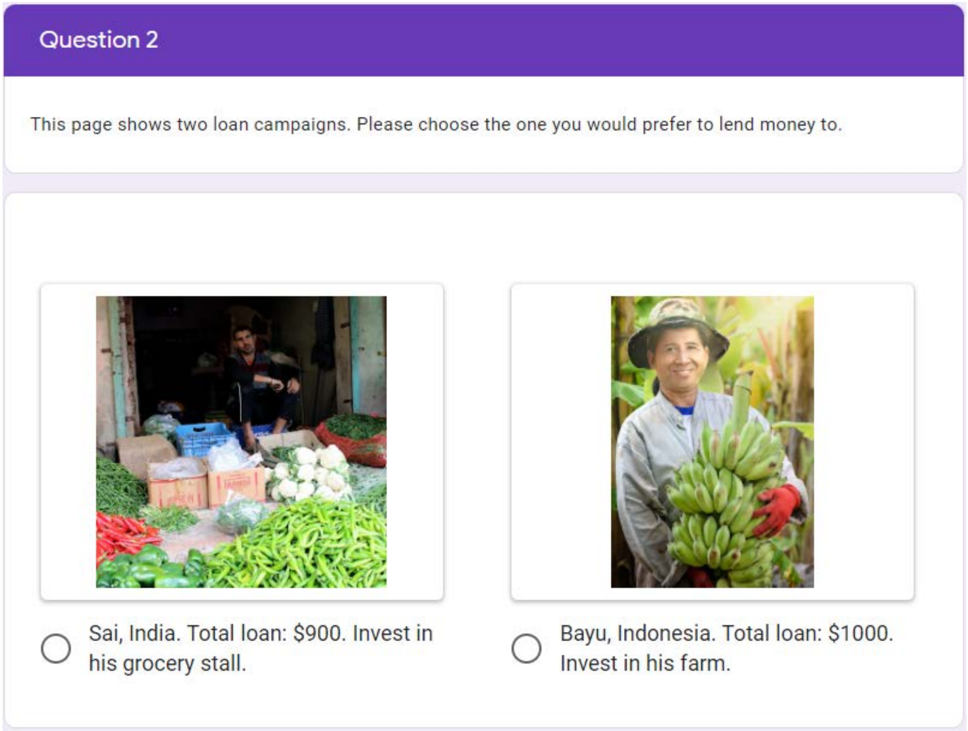


Note: Two versions of an image with variation in male. Both images were generated using GANs.

ilar to real data. GANs have been used in social sciences to generate realistic images (Ludwig and Mullainathan, 2022) and synthetic datasets (Athey et al., 2021). GANs are frequently used to modify images and generate so-called "deep fakes". For our task, we apply Style-GAN developed by Karras et al. (2019). Specifically, we use GANs to vectorize a selected feature of images. Once the feature has been vectorized, we can adjust the vector in the desired direction. The modified attribute is later embedded into the original image while the rest of the image stays unchanged. Finally, we ensure that images look realistic by deblurring, inpainting, and auto-blending. See Appendix B for further discussion and Figures 9 and 10 for examples of GAN generated images.

Experiment implementation. In the experiment, subjects are first introduced to the concept of micro-loans and then presented with six pairs of borrowers and asked to select one in each pair. Figure 11 shows an example of a choice situation. Participation in the experiment took approximately five min-

Figure 11: An example of a choice instance



Question 2

This page shows two loan campaigns. Please choose the one you would prefer to lend money to.

○ Sai, India. Total loan: $900. Invest in his grocery stall.

○ Bayu, Indonesia. Total loan: $1000. Invest in his farm.

Note: An example of choice instance from the recruited experiment. Both images show borrowing campaign profiles featuring males. *The left profile is not a Body-shot and not smiling. The right panel is not a Body-shot and the borrower is smiling.*

utes. The survey included several features that encourage and assess thoughtful responses as detailed in Appendix G.

To generate experimental protocols, we first create a pool of images. Starting with 20 original images, we generated artificial versions with variations in *male*, *smile*, and *body-shot*; thus, each image has 8 versions.[20] All images used in the experiment were artificial, GANs-generated versions of the original images. Therefore, subjects were asked to choose between two fabricated images. In the rest of this section, we use the term ’profile’ when referring to all eight variants derived from the same image.

Second, to allocate images to protocols we draw the first image, without replacement, and pair it with another image such that the version presented differs at least by one feature and the two images were not generated based on the same original image. We continue this until we have six pairs. In total, we created 15 protocols.

[20] Due to privacy and ethical considerations we do not modify images of Kiva borrowers, whom we cannot contact to consent to the modification of their images. Instead, we purchase images from *Shutterstock.com*, a website that sells images. We select images that are similar to images used by Kiva borrowers and use purchased images to train GANs and generate images altered in the desired way.

Sample recruitment. The experiment was carried out on Prolific.co. We recruited 400 subjects that declare to have contributed to a charitable cause in the prior year. We considered subjects from developed countries with high socioeconomic status (self-reported). We impose these criteria to ensure that our subjects are similar to Kiva lenders.

The mean age of subjects in the experiment is 33 years, 51% are female, 49% are male, and 0.001% decided not to provide gender. Subjects were asked to declare the amount they donated to charity in the previous year, we considered eligible for the experiment only those who donated at least USD 1; 60% of respondents donated less than USD 75. We recruited respondents from developed countries. United Kingdom is the most common country of residence of our subjects with 40% of subjects, followed by Spain with 30%, and France with 9%. In Appendix H, we present summary statistics on employment and self-reported socioeconomic status.

Note that the subjects in our experiment are not actual lenders on Kiva. Thus, the preference estimates obtained in our experiment are not reflecting the preferences of Kiva users and should be viewed as additional evidence corroborating the importance of selected features in choosing between microlending campaigns, but not as definitive evidence of their impact on outcomes on the Kiva platform.

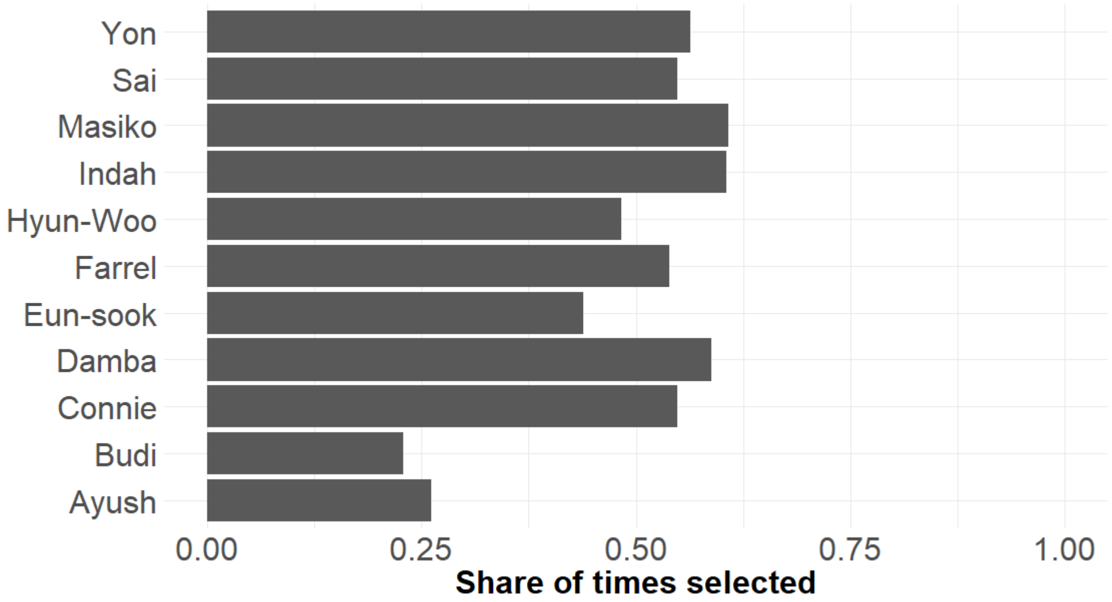
5.2 Experiment results

Subjects in the experiment choose between two profiles in each choice instance. The outcome is whether the profile has been chosen or not, and its mean is 0.5 (there is no option to skip the choice instance). We start by evaluating whether some profiles are on average selected more frequently than others. Figure 12 shows the average outcome per profile (note, each profile had two names male and female, here we present only one of them). We observe that there are clear differences in the popularity of profiles but none of them was selected almost always or almost never. We keep all of them for the analysis (for robustness we will also carry out the analysis without the two least popular profiles).

Average treatment effects. Subjects in the experiment were asked to choose the preferred profile out of a pair. Suppose the preference has a systematic component, which depends on *male*, *smile*, and *Body-shot* and there is also an iid random component ϵ_{ij}. The utility of subject i choosing the option j is written:

$$u_{ij} = \alpha * male_j + \beta * smile_j + \gamma * Bodyshot_j + \mu_j + \epsilon_{ij}. \quad (1)$$

Figure 12: Mean outcomes per profile.



Note: Vertical axis the borrower profile name (we group male and female variants together but show one name). Horizontal axis the number of times the profile has been selected divided by the number of times the profile was shown.

Assuming that ϵ is distributed following a type I extreme value distribution, the probability of a subject i choosing option j, is written:

$$u_{ij} = \frac{\exp(\alpha * male_j + \beta * smile_j + \gamma * bodyshot_j + \mu_j)}{\sum_{k=j,j'} \exp(\alpha * male_k + \beta * smile_k + \gamma * bodyshot_k + \mu_j)}. \tag{2}$$

We are interested in the estimates of parameters $\alpha, \beta, and \gamma$. We obtain them by estimating a logistic regression model by maximizing the conditional likelihood.

Table 5 presents the results, the baseline specification is in column (1), column (2) additionally adjusts for subject-specific covariates, column (3) excludes the least liked profiles (Ayush and Budi), column (4) divides profiles into high and low fixed effects ones and interacts profile features and fixed effects, column (5) add subjets' characteristics to column 4.

We find that all the features of interest are statistically significant and have high magnitudes: *male* and *body-shot* lead to lower outcomes, while *smile* increases outcomes. In column (3) where we exclude low FE profiles *body-shot* is no longer statistically significant. In columns (4) and (5), we divide the profiles into highly attractive and least attractive, based on mean outcomes: two profiles fall into the latter category: Budi, and Ayush. We find that the point estimates of the average treatment effects go in the same direction for both profile types. Average marginal effect of *male* is 31% reduction in the

Table 5: Average treatment effects estimates

	Dependent variable:				
	chosen		chosen	chosen	
	(1)	(2)	(3)	(4)	(5)
male	-0.385^{***} (0.079)	-0.373^{***} (0.080)	-0.747^{***} (0.092)		
smile	0.298^{***} (0.074)	0.331^{***} (0.078)	0.554^{***} (0.088)		
body shot	-0.191^{**} (0.079)	-0.160^{*} (0.084)	-0.121 (0.096)		
male × high FE				-0.372^{***} (0.091)	-0.366^{***} (0.093)
smile × high FE				0.233^{***} (0.081)	0.252^{***} (0.087)
body shot× high FE				-0.176^{**} (0.084)	-0.142 (0.089)
male × low FE				-0.337^{*} (0.195)	-0.318 (0.197)
smile × low FE				0.822^{***} (0.276)	0.944^{***} (0.307)
body shot × low FE				-0.381^{*} (0.226)	-0.414^{*} (0.241)
Image FE	x	x	x	x	x
Subject's characteristics		x	x		x
Restricted sample			x		
Observations	4,920	4,644	4,142	4,920	4,644

Note: Estimates of the logistic regression. Columns (2) and (5) include features of subjects. Columns (4) and (5) divide profiles into high and low fixed effects (low FE are the two least popular profiles - Budi and Ayush). Column (3) restricts the sample to high fixed effects profiles only. **p<0.1;* ***p<0.05;* ****p<0.01*

probability of being selected, 17% drop for the *body-shot*, and 34% increase for *smile*.

To sum up, evidence from the recruited experiment corroborates findings from the observational data. While neither analysis is conclusive together they represent suggestive evidence of the impact of the selected *style* features on outcomes on Kiva.

6 Efficiency - Fairness Tradeoff: counterfactual simulations

Various platform policies can exploit the finding that certain *style* features impact outcomes. In this section, we propose several such policies, simulate counterfactual outcomes, and evaluate their impact on fairness and efficiency. To do that, we consider a simplified model of interactions on Kiva characterized by the parameters from the recruited experiment.

Although we use a stylized approach, our findings can be used to assess what types of policies are likely to do better than others. In practice, our method can be useful in suggesting which policies to prioritize for a randomized experiment.

6.1 A model of a micro-lending platform

Pool of borrowers. The pool of available borrowers is a set of borrowing campaigns that Kiva selects from to decide which of them to display to lenders. The pool of borrowing campaigns can be summarized by a vector of profiles x, where each element is a profile $x_i = (male_i, smile_i, bodyshot_i, \eta_i)$, which describes features of the borrower i; the first element corresponds to the borrower's *type*: *male* or *female*. The two latter features, *smile*, and *body-shot* relate to borrower's *style*; and η_i is a fixed effect which summarizes all other characteristics of the borrower.

The pool of borrowers is exogenously determined and the joint distribution of borrowers' characteristics is denoted as G.

Policy and markets. A market is a set of borrowers shown to a lender. Platform policy $\mathbb{H}$ transforms the joint distribution of borrowers' characteristics from G to H. Specifically, the policy defines $\mathbf{E}_H\left[f|male, \eta\right]$ for $f \in \{smile, bodyshot\}$ the conditional probability of *style* features in the pool of borrowers. Additionally, a policy applies the probability of being shown to lenders $h : (\eta, male, smile, bodyshot) \to [0, 1]$ to the pool of borrowers. Thus, a policy can be summarized as $\mathbb{H} = \{\mathbf{E}_H\left[f|male, \eta\right], \mathbf{h}\}$.

The policies that we consider have two elements: they can impact the distribution of *style* features in the pool of borrowers. This can, for example, be implemented via advice on profile creation, a protocol that requires borrowers to upload several images and selects the most compliant one, or behavioral interventions that nudge borrowers to create compliant profiles. Second, a policy can differentiate the probabilities with which borrowers in the pool appear in the market. We allow the platform to condition these probabilities on borrowers' characteristics.

Lenders. Lenders, indexed by j, arrive to the platform and observe available borrowers. They choose the option that maximizes their utility a borrower or the outside option. The utility associated with choosing one of the borrowers is written:

$$u_{ij} = \alpha_j * male_i + \beta_j * smile_i + \gamma_j * bodyshot_i + \eta_i + \epsilon_{ij}, \tag{3}$$

where $(\alpha_j, \beta_j, \gamma_j)$ are random preference parameters; ϵ_{ij} is a random utility parameter, which is iid across lenders and borrowers, GEV distributed. The utility from choosing the outside option is $u_{oj} = \omega + \epsilon_{oj}$. Lenders choose an option that maximizes their utility.

6.2 Implementation.

Markets. We consider a pool that consists of 22 borrowing campaigns and assume that a market can consist of a maximum of ten borrowers. Campaigns' fixed effects take values of fixed effects estimated in the experiment and the conditional distribution of features G follows the distribution in *Kiva data*.[21] Fixed effects in *Kiva data* are estimated as predicted *cash per day* net of the impact of *male, smile* and *body-shot*. To construct borrowers' profiles, we treat a fixed effect as a random variable $\tilde{\eta}$ drawn from $\mathcal{N}$ a set of fixed effects estimated in the experiment; $\hat{\eta}$ is its realization. Second,

$$\mathbf{E}_G\left[male|D(\hat{\eta})\right] = \mathbf{E}_K\left[male|D(\eta_k)\right],$$

where K stands for distribution in *Kiva data*, $D(\cdot)$ is decile of the fixed effect and η_k is the fixed effect from *Kiva data*. $\mathbf{E}_K\left[male|D(\eta_k)\right]$ is the share of *male* profiles in *Kiva data* per decile; thus, the share of *male* borrowers with the fixed effect in the first decile of fixed effects estimated from the recruited experiment equals the share of *male* borrowers in the lowest decile of *Kiva data* fixed effects. Finally, *style* features are also distributed following the conditional distribution in *Kiva data*:

$$\mathbf{E}_G\left[smile|male, D(\hat{\eta})\right] = \mathbf{E}_K\left[smile|male, D(\eta_k)\right].$$

Thus, we allow the smiling rates to differ across *male* and fixed effects. Analogously *body-shots* are distributed such that:

$$\mathbf{E}_G\left[bodyshot|male, D(\hat{\eta})\right] = \mathbf{E}_K\left[bodyshot|male, D(\eta_k)\right].$$

Lenders preferences. We assume that lenders' preference $(\alpha_j, \beta_j, \gamma_j)$ are parameters drawn from distributions estimated using experimental data, such that $\alpha_j \sim N(\alpha, sd_\alpha)$, where α is the estimate of the average treatment effect and sd_α is its standard error; ϵ_{ij} is a random utility parameter, which is iid across lenders and borrowers, GEV distributed. We set the utility from choosing the outside option to one (the highest FE estimated in the experiment is 0.64).

Outcome metrics. We propose two metrics of fairness: first, to capture overall distribution of outcomes we use the Gini coefficient defined as

$$Gini = \frac{\sum_{j=1}^{n} \sum_{j'=1}^{n} |x_j - x_{j'}|}{2n\bar{x}},$$

[21] Kiva's existing policy is based on the time in which the borrower posted the campaign. Thus, assuming that arrival time is independent of characteristics, a lender sees each borrower in the pool with equal probability. In reality, this is an approximation, because campaigns that reach their funding outcomes are removed from the platform. Thus, the less attractive campaigns stay longer on the platform, so lenders have a higher chance of observing them.

where x_j is the outcome of borrower j and $x_{j'}$ of borrower j', n is the number of borrowers, and $\bar{x}$ the average outcome. Second, to analyze how the outcomes of the worst performing borrowers depend on the platform policy, we use the sum of market shares of borrowers in the bottom tercile. We will compare the outcomes under various policies to a *fair* benchmark, where the distribution of *style* features do not impact the distribution of outcomes.

We measure efficiency as the share of lenders that chose a borrower instead of an outside option. To compute all metrics we consider all borrowers in the pool. Thus, we capture both borrowers that were included in the market as well as those that stayed out.

Market outcomes. To determine market outcomes, we simulate markets and choices by lenders. Based on the distribution of outcomes, we compute fairness and efficiency metrics.

Each simulation proceeds in three steps: first, we simulate the pool of borrowers. To do that we draw 22 fixed effects from $\mathcal{N}$, the pool of fixed effects estimated using data from the experiment, and assign *male* to profiles with the frequency from *Kiva data*. After that, we assign *smile* and *body-shot* following their conditional frequencies.

Second, we construct markets from the pool of borrowers. A policy determines $h(\eta_i, male_i, smile_i, bodyshot_i)$ the probability that a borrower in a pool appears in the market. A market is constructed per lender. This means that in one simulation there is one pool of borrowers, from which borrowers are sampled for each lender.

Finally, we simulate lenders' preferences and their choices as described in Equation 3. We perform 50 simulations of 500 lenders' choices for each policy. We use the outcomes to compute our metrics of fairness and efficiency. We consider all borrowers in the pool, irrespective of whether they were shown to lenders or not. Appendix I presents the algorithm that we used.

6.3 Counterfactual policies

Baseline. Baseline policy represents the existing policy on Kiva. In the baseline policy, each borrower in the pool is assigned an equal probability to be included in the market and the joint distribution of features is G; that is Kiva does influence how *style* features are distributed.

Benchmark: fair. In the *Benchmark* every borrower in a pool has a profile image with a *style* featuring *smile* and without *body-shot*. All borrowers have the same probability of appearing in the market. In the *Benchmark*, we keep the probability of choosing an inside option fixed at the level in *Baseline*. By doing so we can isolate the role played by the distribution of *style* features in shaping the inequity of outcomes and showcase a fairer outcomes distribution.

Naive. In this policy, we show what happens when a platform realizes that profiles with *smile* and without *body-shot* are more attractive and over-samples them. In practice, when the number of borrowers with *smile* and *body-shot* is more than ten, the platform randomly samples from them. Otherwise, the platform includes all compliant borrowers and fills in the empty slots by randomly drawing additional borrowers. In expectation (i.e., before the pool of borrowers is determined), some non-compliant borrowers are always included.

Partial Compliance. In this case, the platform issues a recommendation to everyone to make sure that profile images have *smile* and do not have *body-shot*. In practice, we assume that previously non-compliant borrowers become compliant with a probability of 75%.[22] After the pool is determined, the platform assigns an equal probability of being included in the market to all borrowers.

Low-type support. This policy promotes borrowers that based on their *types* are predicted to have low funding outcomes by ensuring that they are always included in the market. We focus on *gender* in this application. Practically, the approach is analogous to *Naive*; when the number of *male* campaigns is above ten, the platform samples randomly from them. Otherwise, the platform includes all *male* profiles and fills in other slots by randomly selecting from available profiles. In expectation, there are some *female* profiles included in the market.

Restrict Competition. In this policy, the platform promotes fairness by reducing the competition between borrowers. To implement this the platform randomly selects five borrowers from the pool to form the market (instead of ten).

Hybrid. Hybrid policy combines *Partial Compliance* and *Low-type support*.

Note that all the policies that we propose in expectation give non-zero probabilities of being included in the market to any borrower.

[22] Such a profile feature recommendation can be implemented in various ways, for example, through behavioral nudges or a script requiring that several images need to be uploaded from which platform selects the ones to be shown to lenders.

6.4 Results

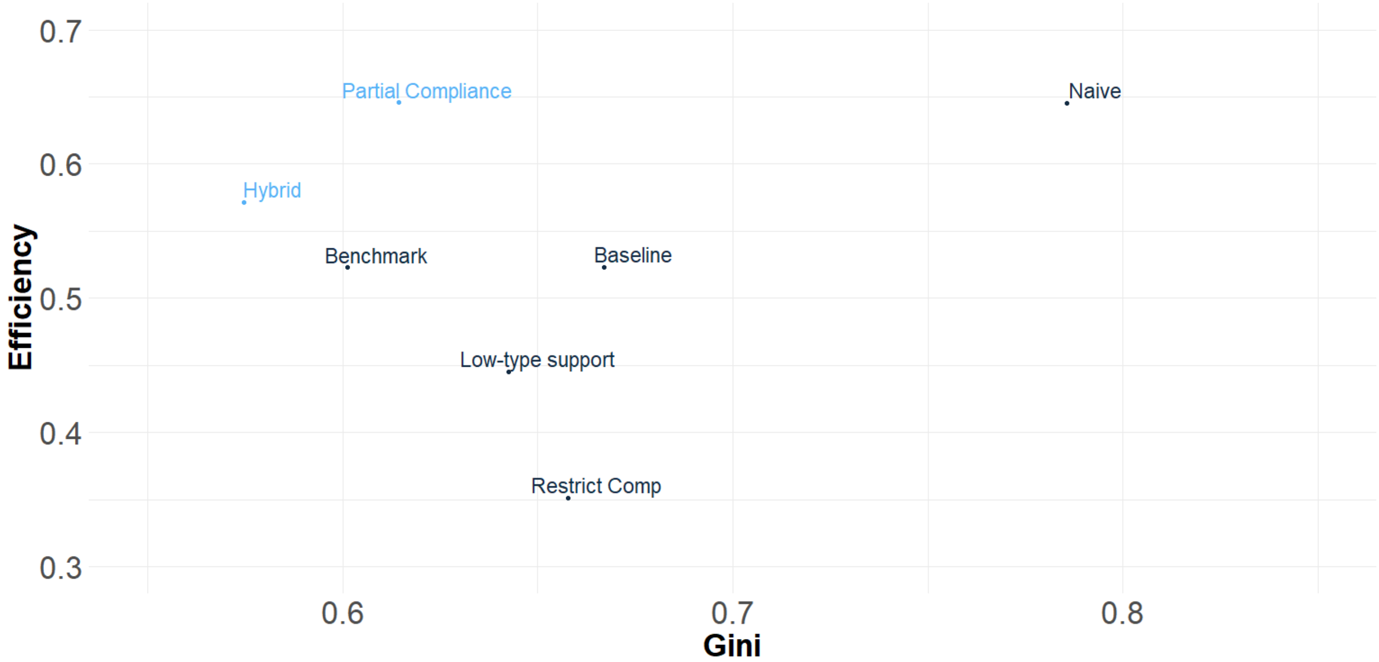
Figure 13 presents the results from simulations of the proposed policies. On the horizontal axis, we show the mean of Gini coefficients from 100 simulations of each policy. On the vertical axis, we show the mean of the shares of lenders choosing a borrower rather than the outside options.

We find that for the parameters that we used, the proposed policies impact both metrics considerably. First, in the *Baseline*, the Gini coefficient is around 0.67 and efficiency 0.54. Second, *Benchmark* showcases the impact of the unequal distribution of *style* features on fairness; when all borrowers have profiles with the desired features the Gini coefficient reduces to 0.58. Next, *Naive* policy has a strong negative impact on fairness. The Gini coefficient increases to almost 0.8. Recall, under this policy the platform includes more profiles with *smile* and *body-shot* in the market. Unfortunately, this policy has the unintended consequence of reducing the prominence of profiles with *types* associated with lower outcomes, further increasing inequities in outcomes' this is due to the correlation between *type*s and *style*. The upside of this policy is that it boosts efficiency as lenders prefer profiles with the selected features. The other policy focused on increasing prominence, *Low-type support*, has exactly the opposite effects. We observe a decrease in efficiency because the platform now includes in the market more borrowers with *types* leading to lower outcomes. As a consequence, this practice leads to more equitable outcomes; because of random preference components of utility, lenders will choose these borrowers more frequently than in the *Baseline*. The alternative pro-fairness policy, *Restrict competition* leads to a small reduction in the Gini coefficient; however, there is a substantial cost to efficiency with fewer options to choose from lenders are more likely to choose the outside option.

Two policies, marked in blue, stand out. *Partial Compliance* leads to gains on both dimensions. On the one side, higher frequencies of *smile* and *body-shot* lead to higher average desirability of the borrowers. On the other hand, due to the lower initial prominence of these features amongst borrowers with *type* features associated with lower outcomes, we can observe a decrease in the Gini coefficient. Finally, *Hybrid* policy combines the effects of *Partial Compliance* and *Low-type support*. There is a substantial reduction in the Gini coefficient of the distribution outcomes and a moderate gain in efficiency.

In Figure 14 we compare the proposed policies using the alternative fairness metric: the sum of market shares of the 33% of the least popular borrowers in the pool. The vertical axis is unchanged - the share of borrowers choosing an inside option. First, in the baseline, only 1% of lenders choose a borrower in the bottom third. Second, all the proposed policies increase the share of the borrowers with the lowest outcomes, except for *Naive* and *Restrict Competition*, where we see a reduction in the

Figure 13: Fairness - Efficiency tradeoff: Gini coefficient.



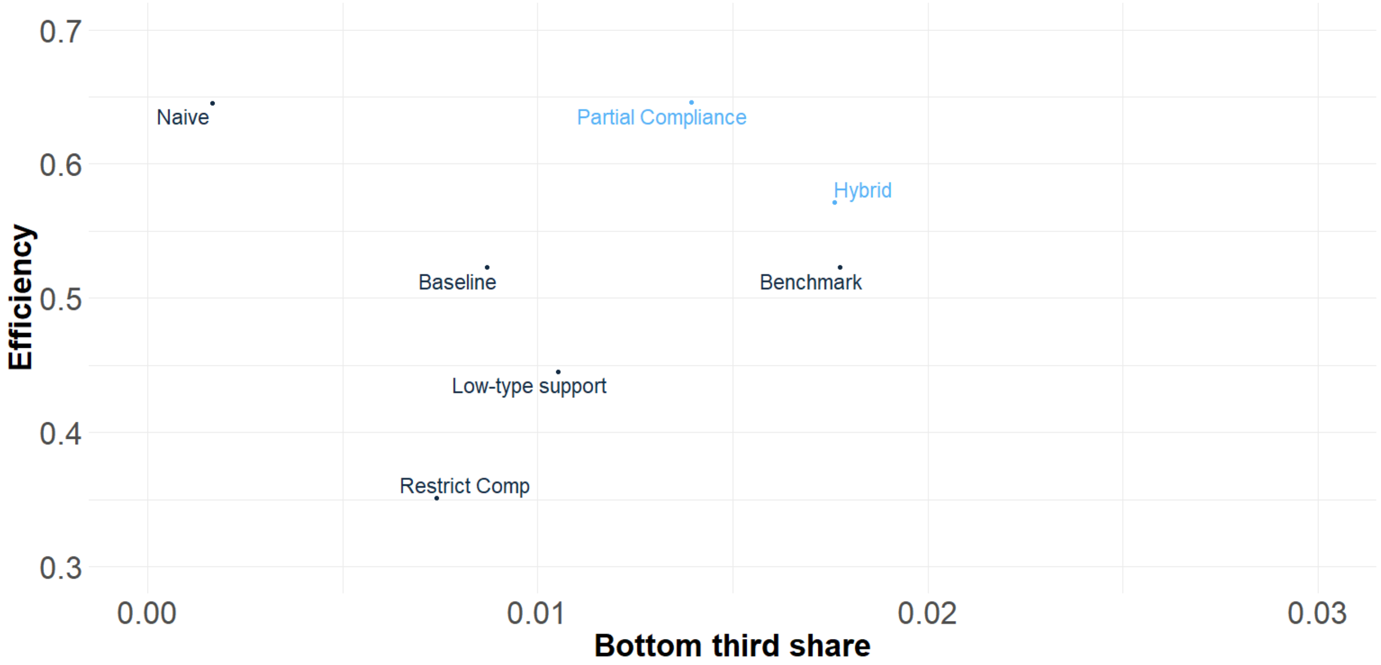
Note: Gini coefficients and efficiency. Each point represents the mean of 100 simulations with 500 lenders each. The horizontal axis presents Gini coefficients while the vertical axis reports the share of lenders choosing an outside option.

market share of the bottom third. *Partial Compliance* policy improves on both dimensions. Finally, *Hybrid* leads to a pronounced increase in the share of the bottom third, up to 2%.

To better visualize the impact of proposed policies, we come back to the histograms presented in section 4.2. In Figure 15, we present histograms of observed and simulated *cash per day*. In red (baseline) we show the distribution in *Kiva data*. In blue we present simulated outcomes.

To obtain simulated outcomes, we, first, randomly draw from a log-normal distribution such that the mean value equals the average *cash per day* in *Kiva data* and variation of the distribution is selected such that the difference in Gini coefficients between baseline and counterfactual is the same, in percentage terms, as in simulations presented in Figure 13. Second, to adjust for the change in efficiency, we increase all values by the percentage difference in efficiency from Figure 13. Both policies move some borrowers from low to moderate outcomes sections of the distribution. However, some high-performing borrowers receive lower outcomes.

Figure 14: Fairness - Efficiency tradeoff: bottom third market share.



Note: Bottom third market share and efficiency. Each point represents the mean of 100 simulations with 500 lenders each. The horizontal axis presents the sum of the market shares of borrowers in the bottom third by outcomes. The vertical axis reports the share of lenders choosing an outside option.

6.5 The impact of proposed policies on *gender* inequity

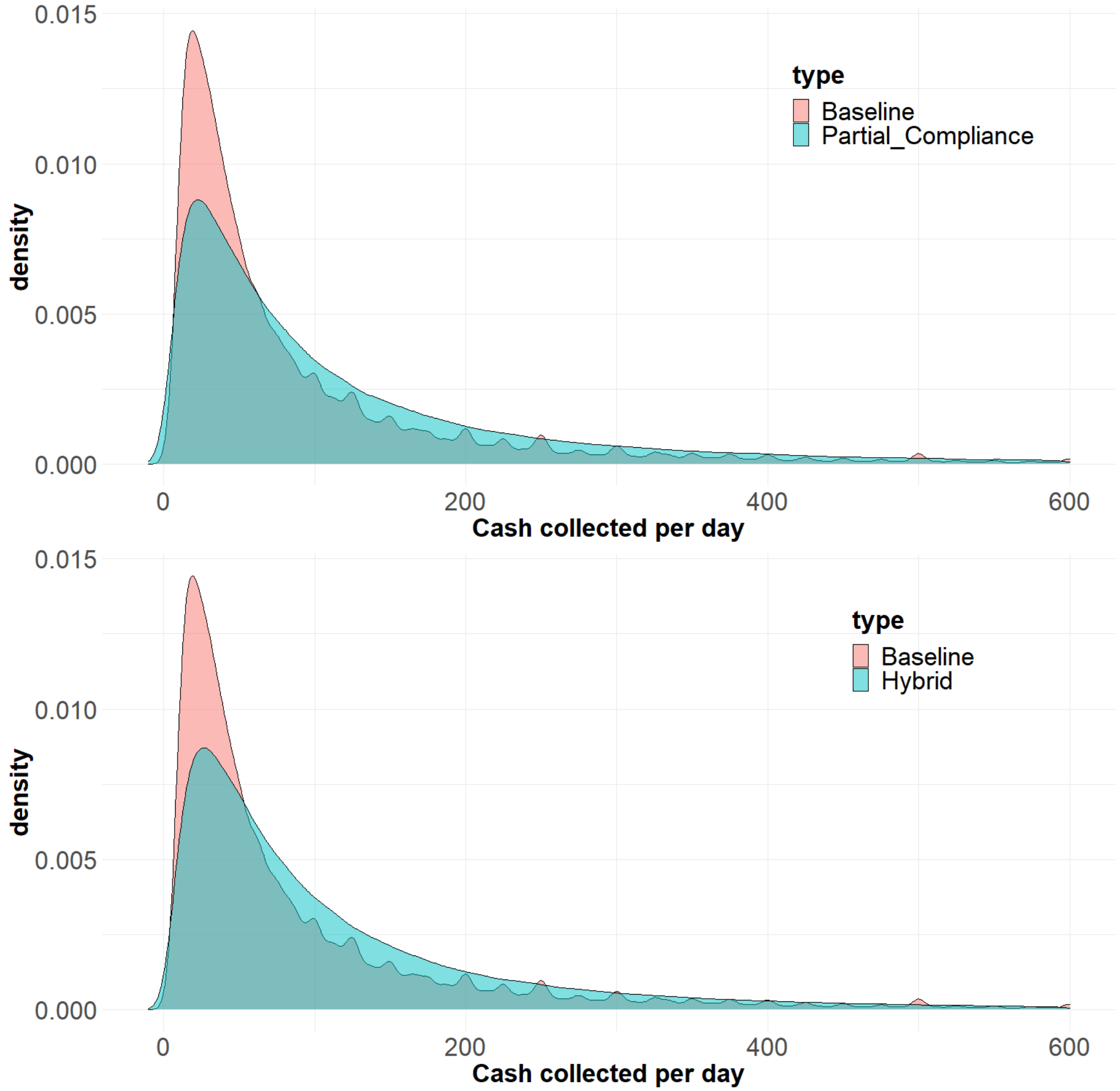
In section 4.5, we documented a substantial disparity in outcomes across *genders*.[23] In Figure 16, we present the impact of proposed policies on *gender* equity and efficiency. In the *Baseline* only around 25% of lenders choose a borrower with a *male* profile (we adjust the share of selected *male* borrowers by the share of *male* borrowers in the pool). A part of the gender gap can be associated with different frequencies of the selected *style* features. From *Benchmark* we can observe that when all borrowers have profiles with *smile* and without *body-shot*, the share of lenders choosing a borrower with a *male* profile increases to approximately 30%.

All discussed policies, except for *Restrict Competition*, increase the share of lenders choosing *male* borrower. However, we observe that *Low-type support* and *Hybrid* policies result in the disparity going in the opposite direction. Now, campaigns of *male* type substantially outperform *female* campaigns. The *Hybrid* policy boosts the share of lenders choosing a borrower with a *male* profile to above 84% and leads to a moderate gain in efficiency. Finally, *Partial compliance* increases both fairness and efficiency.

The results of the counterfactual simulations confirm the logic described earlier: when desirable

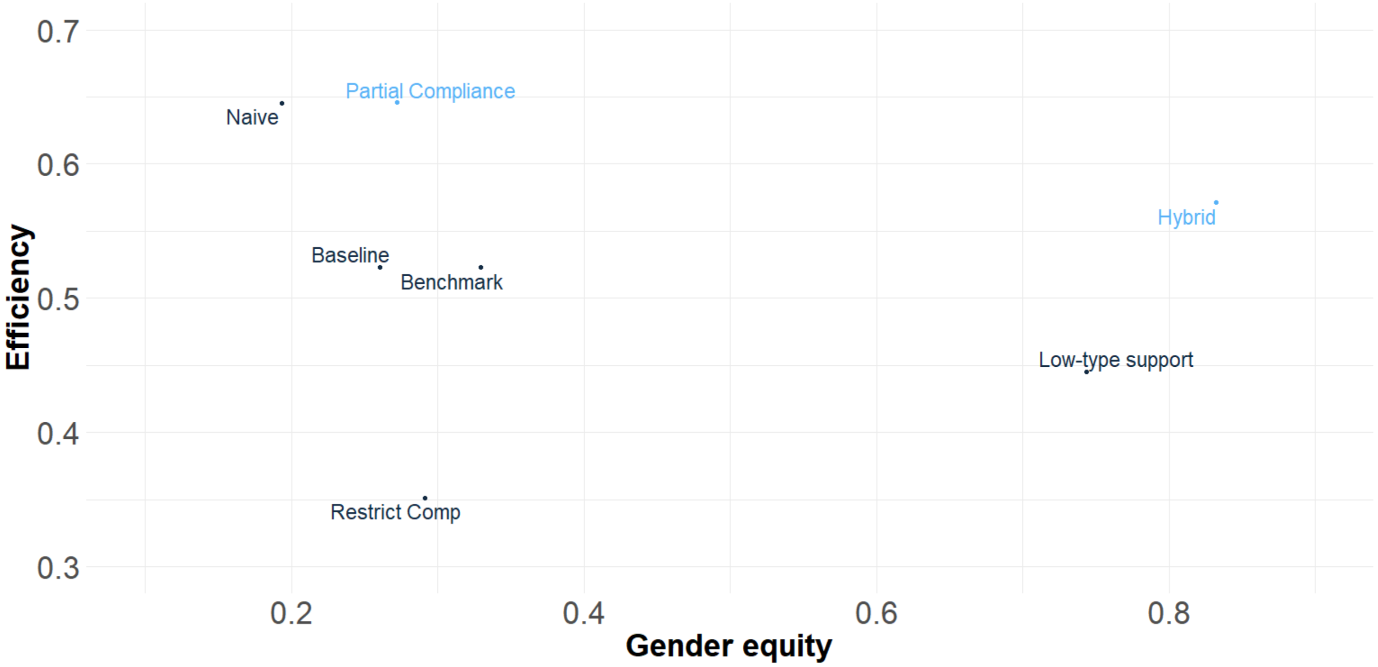
[23] Recall that we use the algorithmic prediction of gender; thus, the inequity that we consider relates to the disparity between profiles that the algorithm classifies as *male* and *female*.

Figure 15: Histograms of distribution of cash per day under baseline and selected counterfactual policies.



Note: Histograms of cash collected per day; Baseline in red: observed in Kiva data, in blue simulated from counterfactual policies. Partial Compliance *on the top panel and* Hybrid *on the bottom.) Values for the counterfactual policies simulated from log-normal distribution to match the values in Figure 13.*

Figure 16: Fairness - Efficiency tradeoff: gender disparity.



Note: Gender disparity and efficiency. Each point represents the mean of 50 simulations with 1000 lenders each. The horizontal axis presents the share of lenders choosing a borrower with a male profile adjusted for the share of male profiles in the borrower pool. The vertical axis reports the share of lenders choosing an outside option.

style features are positively correlated with *type* characteristics leading to better funding outcomes, platform policies that promote the selected *style* characteristics by increasing the prominence of profiles with them aggravate inequities in outcomes. In contrast, policies that change the distribution of attractive *style* features in a way that increases their prevalence amongst borrowers with high *type* characteristics lead to a more equitable distribution of outcomes. Furthermore, policies that increase the overall share of profiles with desirable *style* also boost efficiency.

The proposed model is a simplification of interactions between lenders, borrowers, and the microlending platform. Amongst other assumptions, we abstract from supply side responses or the possibility of lenders choosing several campaigns. Therefore, the specific magnitudes of the impact of the discussed policies on fairness and efficiency should be measured in a randomized experiment. Nevertheless, the proposed approach provides strong support for testing policies based on *style* recommendation.

7 Conclusion

In this paper, we consider a problem of an online marketplace that wants to balance fairness and efficiency in the context in which users have preferences for features of profile images. We introduce a distinction between *type* and *style* features: the former characteristics are fixed when users create their online profiles, and the latter ones are determined during the profile creation.

Using observational data from a large microfinance platform, we, first, show high inequities in funding outcomes, second, we demonstrate that *style* features are predictive of funding outcomes but not of defaults, and third, we showcase selected *style* features that have an impact on funding outcomes and are correlated with borrowers' *types*; in consequence, they exacerbate outcomes inequity in a way that, we argue, is unfair to borrowers. To corroborate these findings, we carry out a recruited experiment. In the experiment, we use Generative Adversarial Networks to generate fabricated images with a variation in selected features. We document that subjects prefer profiles with *smile* that are not *body-shots*. Finally, we counterfactually evaluate various platform policies exploiting the estimates of the impact of selected *style* features on outcomes and their correlation with *types*. We show that a policy that encourages profiles with desirable features increases fairness and leads to more transactions.

The mechanism underlying our findings is that unchangeable *type*s and amenable *style*s can both lead to inequities, and their correlation determines which platform policies will be effective in promoting fairness without sacrificing efficiency. In the case of a positive correlation between desirable *style* features and high *type*s, platform policies which promote profiles with selected features are likely to lead to less fair outcomes. In contrast, policies that shift distributions of these features in the direction that increases their adoption by low-performing users can promote fairness.

Our approach can be used to determine which policy classes have the potential to increase fairness and efficiency. However, the extent to which the proposed policies are effective depends on specific parameters of lenders' demand (magnitudes and their stability) and borrowers' responsiveness to recommendations. Therefore, we believe that the pipeline we propose can be useful to prioritize policies for a randomized experiment. Carrying out such an experiment is a natural next step in this research agenda.

Throughout the paper, we abstracted from the developmental impact of borrowing campaigns and argued that outcomes inequities which are not justified by different repayment probabilities and other covariates are unfair. We did not adjust for developmental impact due to the lack of appropriate measures; doing so is a valuable extension of this research.

References

Abbey, J. D. and Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management*, 53:63–70.

Aggarwal, R., Goodell, J. W., and Selleck, L. J. (2015). Lending to women in microfinance: Role of social trust. *International Business Review*, 24(1):55–65.

Ai, W., Chen, R., Chen, Y., Mei, Q., and Phillips, W. (2016). Recommending teams promotes prosocial lending in online microfinance. *Proceedings of the National Academy of Sciences*, 113(52):14944–14948.

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*, 11:45–66.

Ash, E., Durante, R., Grebenshchikova, M., and Schwarz, C. (2022). Visual representation and stereotypes in news media.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. (2021). Using Wasserstein Generative Adversarial Networks for the Design of Monte Carlo Simulations. *Journal of Econometrics*.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Banerjee, A., Karlan, D., and Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, 7(1):1–21.

Brock, J. M. and De Haas, R. (2021). Discriminatory lending: Evidence from bankers in the lab. *CentER Discussion Paper*.

Duarte, J., Siegel, S., and Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *The Review of Financial Studies*, 25(8):2455–2484.

D'Espallier, B., Guérin, I., and Mersland, R. (2011). Women and repayment in microfinance: A global analysis. *World Development*, 39(5):758–772.

Edelman, B., Luca, M., and Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American economic journal: applied economics*, 9(2):1–22.

Edelman, B. G. and Luca, M. (2014). Digital discrimination: The case of Airbnb. com. *Harvard Business School NOM Unit Working Paper*, (14-054).

Ert, E., Fleischer, A., and Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism management*, 55:62–73.

Flores-Macías, G. and Zarkin, J. (2022). Militarization and perceptions of law enforcement in the developing world: Evidence from a conjoint experiment in mexico. *British Journal of Political Science*, 52(3):1377–1397.

Fong, C. M. and Luttmer, E. F. (2009). What determines giving to Hurricane Katrina victims? Experimental evidence on racial group loyalty. *American Economic Journal: Applied Economics*, 1(2):64–87.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Galak, J., Small, D., and Stephen, A. T. (2011). Microfinance decision making: A field study of prosocial lending. *Journal of Marketing Research*, 48(SPL):S130–S137.

Ge, Y., Knittel, C. R., MacKenzie, D., and Zoepf, S. (2016). Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research.

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics*, 34(2):509–543.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1):1–30.

Jenq, C., Pan, J., and Theseira, W. (2015). Beauty, weight, and skin color in charitable giving. *Journal of Economic Behavior & Organization*, 119:234–253.

Johannemann, J., Hadad, V., Athey, S., and Wager, S. (2019). Sufficient representations for categorical variables. *arXiv preprint arXiv:1908.09874*.

Jordan, S. R., Rudeen, S., Hu, D., Diotalevi, J. L., Brown, F. I., Miskovic, P., Yang, H., Colonna, M., and Draper, D. (2019). The difference a smile makes: Effective use of imagery by children's nonprofit organizations. *Journal of Nonprofit & Public Sector Marketing*, 31(3):227–248.

Karlan, D. and Morduch, J. (2009). Access to Finance. In Rodrick, D. and Rosenzweig, M. R., editors, *Handbook of Development Economics*, volume 5. Elsevier.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Kasy, M. and Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 576–586.

Kline, P. M., Rose, E. K., and Walters, C. R. (2021). Systemic discrimination among large us employers. Technical report, National Bureau of Economic Research.

Krumhuber, E., Manstead, A. S., Cosker, D., Marshall, D., Rosin, P. L., and Kappas, A. (2007). Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, 7(4):730.

Kung, F. Y., Kwok, N., and Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology*, 67(2):264–283.

Landry, C. E., Lange, A., List, J. A., Price, M. K., and Rupp, N. G. (2006). Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment. *The Quarterly Journal of Economics*, 121(2):747–782.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., and Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes. *Philosophy & Technology*, 31(4):611–627.

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113(521):390–400.

Ludwig, J. and Mullainathan, S. (2022). Algorithmic Behavioral Science: Machine Learning as a Tool for Scientific Discovery. *Chicago Booth Research Paper*, (22-15).

Mendes, W. B. and Koslov, K. (2013). Brittle smiles: positive biases toward stigmatized and outgroup targets. *Journal of Experimental Psychology: General*, 142(3):923.

Mullainathan, S., Noeth, M., and Schoar, A. (2012). The market for financial advice: An audit study. Technical report, National Bureau of Economic Research.

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., et al. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, 18(1):36–88.

Park, J., Kim, K., and Hong, Y.-Y. (2019). Beauty, gender, and online charitable giving. *Available at SSRN 3405823*.

Pham, C. and Septianto, F. (2019). A smile–the key to everybody's heart? the interactive effects of image and message in increasing charitable behavior. *European Journal of Marketing*.

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):53–92.

Ravina, E. (2019). Love & loans: The effect of beauty and personal characteristics in credit markets. *Available at SSRN 1107307*.

Rhue, L. and Clark, J. (2020). Automatically Signaling Quality? A Study of the Fairness-Economic Tradeoffs in Reducing Bias through AI/ML on Digital Platforms. *Working Paper, NYU Stern School of Business*.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Salminen, J., Şengün, S., Santos, J. M., Jung, S.-G., and Jansen, B. (2022). Can Unhappy Pictures Enhance the Effect of Personas? A User Experiment. *ACM Transactions on Computer-Human Interaction*, 29(2):1–59.

Septianto, F. and Paramita, W. (2021). Sad but smiling? how the combination of happy victim images and sad message appeals increase prosocial behavior. *Marketing Letters*, 32(1):91–110.

Stigler, M. (2018). dec_covar: R implementation of Gelbach covariate decomposition. *https://github.com/MatthieuStigler/Misconometrics/tree/master/Gelbach_decompo*.

Sun, L., Kraut, R. E., and Yang, D. (2019). Multi-level modeling of social roles in online micro-lending platforms. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–25.

Theseira, W. (2009). *Competition to default: Racial discrimination in the market for online peer-to-peer lending*. PhD thesis, Dissertation, Wharton.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Williams, B. A., Brooks, C. F., and Shmargad, Y. (2018). How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *Journal of Information Policy*, 8(1):78–115.

Younkin, P. and Kuppuswamy, V. (2018). The colorblind crowd? Founder race and performance in crowdfunding. *Management Science*, 64(7):3269–3287.

Zhang, S., Mehta, N., Singh, P. V., and Srinivasan, K. (2021). Can an AI algorithm mitigate racial economic inequality? An analysis in the context of Airbnb. *Working Paper*.

Zhang, S. and Yang, Y. (2021). The unintended consequences of raising awareness: Knowing about the existence of algorithmic racial bias widens racial inequality. *Available at SSRN*.

Appendix

A Feature Detection Algorithms

Mask-RCNN. To structurally obtain features of images we use Mask-RCNN. The Mask-RCNN algorithm, developed by Facebook, detects objects from images. As shown in Figure 17, It takes in an image in the input layer and returns an estimated "package" for each object, including the class name, the bounding box, and the mask of each object detected, and those predictions are jointly optimized through the loss function.

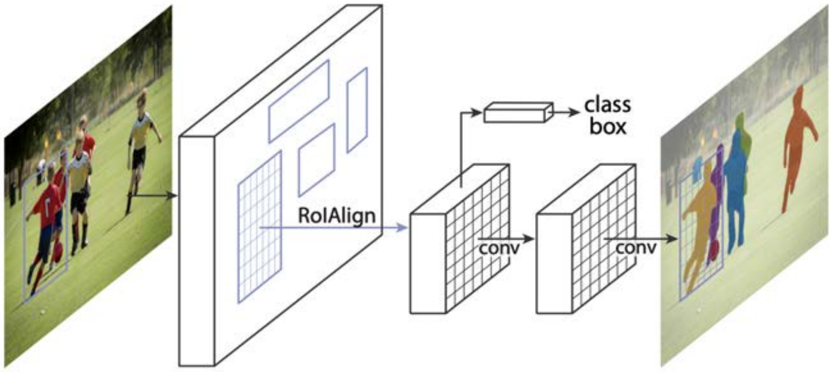


Figure 17: The Mask R-CNN framework [24]

Object detection. We apply this pre-trained model and estimate a score for each object which varies from 0 to 1. The score represents the algorithm's confidence in the existence of a specific feature, such as a tree, person, animal, digital items, etc. Figure 18 visualizes the output. We also apply this algorithm to detect human body-shot. [25]

Facial feature classification. We detect facial features using the *face-classification* algorithm that takes in one face image and outputs a face embedding vector, evaluated by a pre-trained neural network.[27] Then, the embedding vector, as well as the feature labels, enter another neural network model (Multi-layer Perceptron). This model takes in one facial embedding vector and assigns a score for each unique facial feature such as *race*, *gender*, *smile*, etc. It is a supervised learning process, and the training label is pre-annotated.

The features that we obtain from images can be informally classified into three categories: (i)

[25] https://github.com/facebookresearch/detectron2

[27] https://github.com/wondonghyeon/face-classification

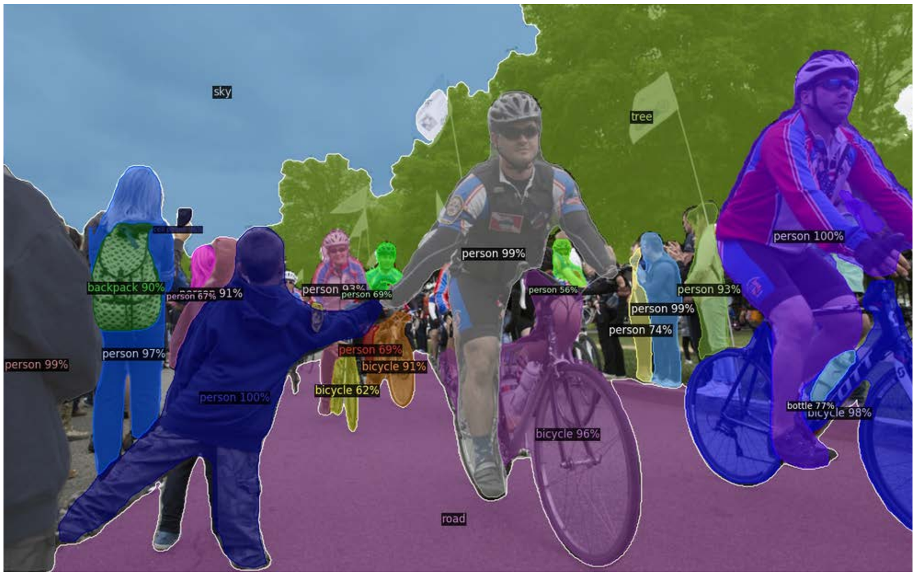


Figure 18: An example outcome of image detection using Mask-RCNN. Each detected object was given a label, put on a mask, and given the corresponding probability score.[26]

technical aspects of the image (e.g., *blurry*, *flash*, *harsh light*), (ii) personal characteristics (e.g., *straight hair*, *eyes open*, *pale skin*), (iii) objects in the image (e.g., *chair*, *clock*).

Technical aspects of the image and personal characteristics (races, ages, hair color, facial shape, eyes/nose characteristics) are detected by FaceNet model which was pre-trained and tested on the large dataset CelebA with over 200,000 facial images. The algorithm identifies the face and detects its features. Images of faces have fixed landmarks and key points. The face detection algorithm, if pre-trained under a large training dataset, can normally attain good accuracy. The pre-training dataset is rich enough to cover all potential variances of Kiva images.

We evaluated specifically the output, *race*, using country (continent) information given by Kiva. We manually compare the predicted race and its country information, and ensure that our CNN algorithm has a good prediction.

B Generative Adversarial Networks

The algorithm described above detects our features of interest. To modify images with respect to these features we use another tool known as Generative Adversarial Networks (GANs). GANs designed by Goodfellow et al. (2014) are an approach to generative modeling using deep learning methods. The key objective of GANs is to generate fabricated data that are similar to particular data, such as realistic images (Ludwig and Mullainathan, 2022) and synthetic datasets (Athey et al., 2021). GANs, although do not directly produce estimates of the density or distribution function at a particular point, can be thought of as implicitly estimating the distribution of latent features, and they can be used to generate or output new examples that plausibly could have been drawn from the original dataset.

The core idea of GAN is to have two models: a generator G and a discriminator D. As illustrated by Goodfellow et al. (2014), to learn the generator's (image) distribution p_g over data x, we define a prior on input noise variables $p_z(z)$, then represent a mapping to data space as $G(z;\theta_g)$. Discriminator $D(x;\theta_d)$ outputs a single scalar, representing the probability that x came from the data rather than p_g. D and G play the two-player minimax with the value function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[logD(x)] + \mathbb{E}_{z\sim p_z(z)}[1 - logD(G(z))]$$

GANs are frequently used to modify images and generate so-called "deep fakes" - fabricated images based on input images that have been altered in a specific way. In our work, we apply Style-GAN developed by Karras et al. (2019) to generate fake images that differ in a specific feature. Conditioning on the attributes (areas) we make the change, the algorithm detects the key image area to leave its counterpart unchanged.

The key image is fed into a pre-trained GAN generator and embedded, as a latent vector V, into a latent space. We compute the direction of the gradient ∇V of our feature of interest W (e.g. smiling), determined from our ATE analysis, by computing the "difference in means" of the latent vector encoded into the latent space from images with and without such feature[28].

$$\nabla V = \mathbb{E}[V_i[W_i = 1, X_i = x]] - \mathbb{E}[V_i[W_i = 0, X_i = x]]$$

We have hyper-parameters to decide the extent to which we want to alter the images in the desired direction. We fine-tune the hyper-parameters image by image to offset the correlation in image features bleeding into the pre-trained GAN model. The modified attribute is embedded into its unchanged counterpart, and we ensure that images look realistic by deblurring, inpainting, and auto-blending.

C Summary statistics of *Kiva data*

[28] We used around 200 images labeled by CNN and verified by human audit

Table 6: Summary statistics of *Kiva data*

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Loan amount	420,765	800.107	993.370	25	275	950	50,000
Cash per day	420,765	123.522	270.186	1	25	116.7	8,750
Days to raise	420,765	13.427	11.667	1	5	20	83
Total number of lenders	420,765	0.012	0.015	0.001	0.005	0.015	0.967
default	420,765	0.050	0.218	0	0	0	1
Male	420,765	0.198	0.398	0	0	0	1
Number of borrowers	420,765	1.958	3.171	1	1	1	50
No. competitors	420,765	0.091	0.173	0.003	0.006	0.075	1.000
Same race gender share	420,765	0.665	0.294	0	0.4	1	1
Asian	420,765	0.191	0.261	0.0001	0.016	0.266	0.995
White	420,765	0.218	0.265	0.001	0.031	0.323	0.999
Black	420,765	0.167	0.281	0.0001	0.006	0.148	0.990
Baby	420,765	0.004	0.003	0.0001	0.002	0.006	0.067
Child	420,765	0.073	0.056	0.001	0.034	0.095	0.609
Youth	420,765	0.264	0.211	0.0002	0.092	0.391	0.982
Middle.Aged	420,765	0.084	0.093	0.0004	0.026	0.104	0.898
Senior	420,765	0.041	0.079	0.0001	0.004	0.039	0.950
Black.Hair	420,765	0.388	0.242	0.0005	0.171	0.589	0.970
Blond.Hair	420,765	0.007	0.029	0.00000	0.001	0.004	0.943
Brown.Hair	420,765	0.405	0.156	0.012	0.288	0.517	0.919
Bald	420,765	0.037	0.073	0.0001	0.004	0.030	0.835
No.Eyewear	420,765	0.865	0.148	0.007	0.830	0.959	1.000
Sunglasses	420,765	0.017	0.014	0.001	0.008	0.020	0.327
Mustache	420,765	0.072	0.160	0.00003	0.004	0.046	0.998
Smiling	420,765	0.549	0.177	0.013	0.424	0.685	0.966
Chubby	420,765	0.339	0.190	0.012	0.185	0.466	0.972
Blurry	420,765	0.162	0.095	0.006	0.090	0.214	0.758
Harsh.Lighting	420,765	0.339	0.165	0.031	0.217	0.430	0.930
Flash	420,765	0.245	0.126	0.010	0.148	0.322	0.855
Soft.Lighting	420,765	0.677	0.090	0.222	0.623	0.742	0.943
Outdoor	420,765	0.447	0.140	0.045	0.343	0.545	0.914
Curly.Hair	420,765	0.394	0.155	0.031	0.275	0.499	0.932
Wavy.Hair	420,765	0.226	0.170	0.004	0.095	0.312	0.991
Straight.Hair	420,765	0.606	0.178	0.034	0.489	0.741	0.982
Receding.Hairline	420,765	0.205	0.235	0.0004	0.039	0.282	0.995
Bangs	420,765	0.171	0.171	0.001	0.052	0.229	0.993
Sideburns	420,765	0.145	0.195	0.001	0.025	0.168	0.977
Partially.Visible.Forehead	420,765	0.094	0.090	0.001	0.032	0.125	0.834
Arched.Eyebrows	420,765	0.451	0.213	0.004	0.282	0.618	0.978
Narrow.Eyes	420,765	0.588	0.204	0.031	0.432	0.755	0.992
Eyes.Open	420,765	0.871	0.073	0.338	0.834	0.925	0.991
Big.Nose	420,765	0.730	0.190	0.042	0.606	0.886	0.998
Big.Lips	420,765	0.586	0.215	0.014	0.425	0.766	0.986
Mouth.Closed	420,765	0.303	0.146	0.018	0.193	0.390	0.944
Mouth.Wide.Open	420,765	0.057	0.040	0.002	0.030	0.072	0.516
Square.Face	420,765	0.019	0.041	0.00005	0.002	0.015	0.759
Round.Face	420,765	0.201	0.155	0.002	0.078	0.287	0.908
Color.Photo	420,765	0.948	0.026	0.632	0.935	0.966	0.997
Posed.Photo	420,765	0.486	0.132	0.069	0.391	0.581	0.925
Attractive.Woman	420,765	0.125	0.151	0.001	0.028	0.158	0.989
Indian	420,765	0.061	0.098	0.00002	0.009	0.066	0.962
Bags.Under.Eyes	420,765	0.586	0.170	0.016	0.468	0.717	0.967
Rosy.Cheeks	420,765	0.122	0.069	0.011	0.072	0.155	0.729
Shiny.Skin	420,765	0.215	0.121	0.004	0.121	0.288	0.808
Pale.Skin	420,765	0.334	0.171	0.014	0.192	0.460	0.908
Strong.Nose.Mouth.Lines	420,765	0.611	0.172	0.026	0.496	0.746	0.966
Flushed.Face	420,765	0.102	0.050	0.009	0.067	0.126	0.573
Top	420,765	157.544	106.715	0	80	204	1,598
Right	420,765	410.062	174.165	29	271	534	960
Bottle	420,765	0.503	2.259	0	0	0	99
Chair	420,765	0.125	0.498	0	0	0	24
Person	420,765	2.119	3.002	1	1	2	39
Bodyshot	420,765	0.406	0.491	0	0	1	1

D Choice of the predictive model

In this section, we consider several predictive models over three specifications and determine the model to be used in the baseline analysis.

We analyze the performance of models predicting *cash per day*. We consider the following models: Linear Regression, LASSO, Random Forrest (grf), and Boosted Random Forrest (grf and gbm). All models (except for LM) are tuned for the task at hand, we report the performance of the selected best (lowest MSE) model. All models are trained using a 70% sample of *Kiva data* and tested on the 30%.

We consider three specifications differing by the number of covariates: (A) covariates include: details of the loan including amount, repayment scheme, *sector*, *country*, etc. and weekly dummies, (B) details of the photo including both *type* and *style* characteristics, (C) total number of active lenders in this *week*sector*, total number of competitors in this *week*sector*, number of competitors of the same *race* and *gender*, and interaction of *week* and *sector*, and interaction of *week* and *country*. For boosted Forrest we also add a 4th specification where we have a sufficient representation of *week* sector* (D) (Johannemann et al., 2019). Table 7 presents results.

Table 7: Comparison of the test-set predictive performance of selected model

s

Model	Specification	MSE	SE
Linear regression	A	13840	159
Linear regression	B	13466	155
Linear regression	C	13565	166
LASSO	A	13797	161
LASSO	B	13379	157
LASSO	C	13183	156
Random forest	A	13930	163
Random forest	B	13530	145
Random forest	C	13099	157
Boosted forest (gbm)	A	12235	156
Boosted forest (gbm)	B	11477	141
Boosted forest (gbm)	C	10929	157
Boosted forest (gbm)	D	11406	173
Boosted forest (grf)	A	12665	147
Boosted forest (grf)	B	12003	149
Boosted forest (grf)	C	11777	139
Boosted forest (grf)	D	11962	177

Note: Test set performance of selected predictive models with different sets of covariates.

We conclude that Boosted Forrest has the best test-set predictive performance across all specifications and we decide to use it as a baseline model for the predictive tasks throughout the paper. *GBM* implementation of the Boosted Forrest has better performance than *GRF*, the difference is moderately small. Sufficient representation does not improve models' performance and will not be used in the predictive tasks.

E Analysis of defaults across default types

We observe two different reasons for the loan not being repaid: a default by a microfinance organization and a default by the borrower. It's plausible that image features are predictive of a borrower's default but not of the microfinance organization. In this section, we separately analyze the predictive performance of a model trained to predict defaults by the borrower with and without image features.

We train a Boosted Forrest (GBM) on 70% of data and report the predictive performance on the 30% test set. We consider two specifications a full model, model C in Dwith defaults as the dependent variable, and a model from which we remove image features. Table 8 reports results.

Table 8: Comparison of the test-set predictive performance of models of default with and without image features.

Outcome	Covariates	MSE	Std. error
All defaults	full model	0.059	0.00094
All defaults	no image covariates	0.059	0.00095
Defaults by the borrower	full model	0.046	0.00088
Defaults by the borrower	no image covariates	0.046	0.00088

Note: Test set performance of selected predictive models with different sets of covariates.

These results suggest that image characteristics do not improve the predictive performance of either of the default models.

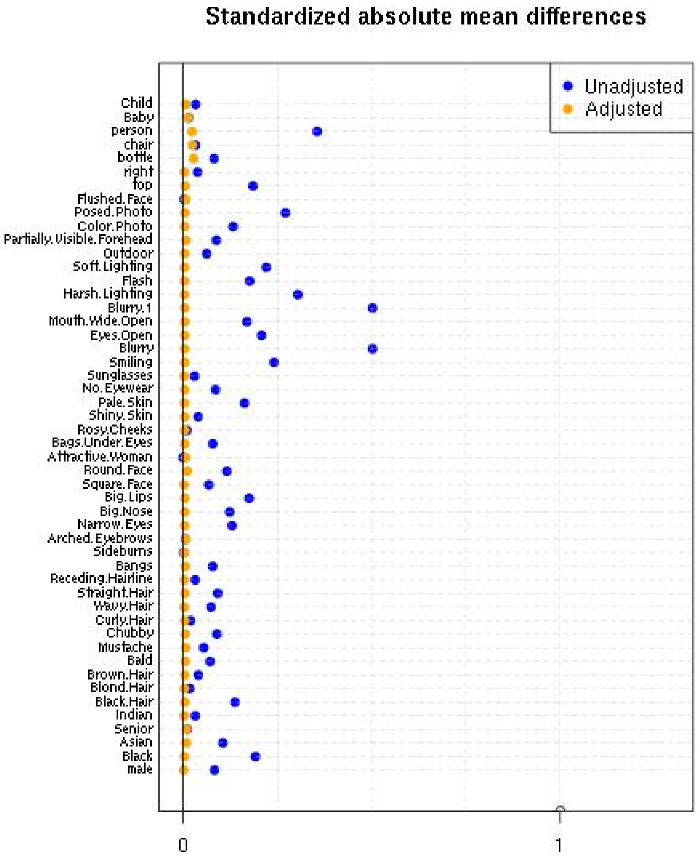
F Supplementary analysis for AIPW estimates of style features

F.1 Diagnostics for selected *style* features

Diagnostics *bodyshot*. Figure 19 shows standardized absolute mean differences of covariates across treatment group (with *Bodyshot*) and control. We see that the adjusted values (yellow) are well balanced.

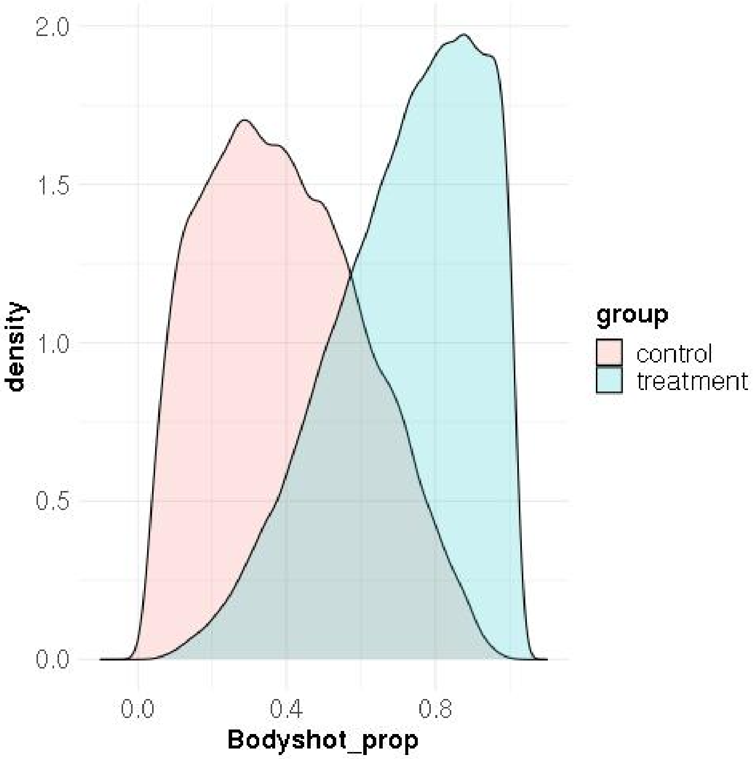
In Figure 20 we show the propensity scores to have a profile with *Bodyshot* across profile images with and without a *Bodyshot*. We find that there is common support between the two groups.

Figure 19: Diagnostics for *bodyshot*



Note: Standardized absolute mean differences of a selected subset of other covariates across profiles with and without bodyshot. *Propensity score used for reweighing obtained using GBM model trained on all covariates in* Kiva data.

Figure 20: Propensity scores for *Bodyshot*



Note: Estimates of the propensity for an image to be a Bodyshot *using a GBM-based prediction with full set of controls from* Kiva data.

Diagnostics *smile*. Figure 21 shows standardized absolute mean differences of covariates across the treatment group (with *smile*) and control. We see that the adjusted values (yellow) are well balanced.

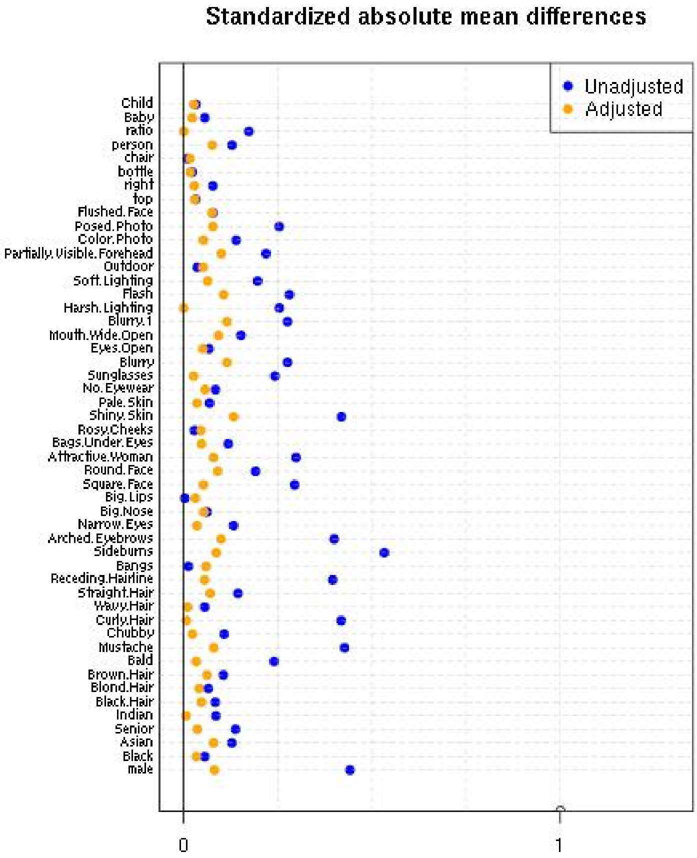
In Figure 22 we show the propensity scores to have a profile with *smile* across profile images with and without a *smile*. We find that there is common support between the two groups.

F.2 ATE estimates using alternative models

Table 9 presents AIPW estimates of the selected *style* features on *cash per day* using additional models. The objective is to show that the findings presented in 5 are robust to the choice of model . Columns two and three of Table 9 show results based on the regression forests model for both the outcome and propensity, with observations reweighted using Li et al. (2018) method. Columns four and five are based on GBM predictions of the outcome and propensity model, and columns six and seven are also based on GBM but with additionally reweighing as in Li et al. (2018).

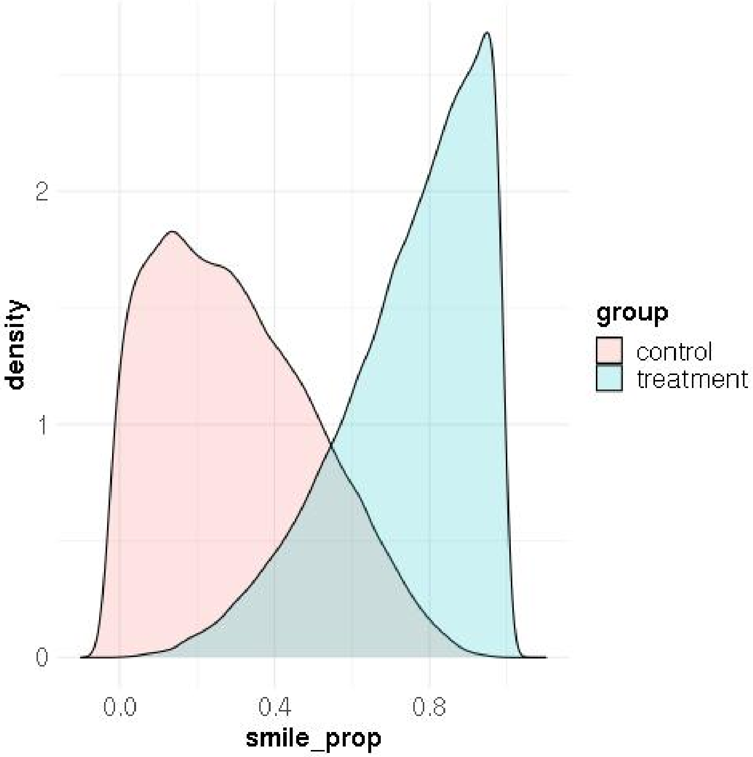
We conclude that while the point estimates change with the choice of the model, the main conclusion that *smile* has a positive impact on *cash per day* and *body-shot* has a negative is robust.

Figure 21: Diagnostics for *Bodyshot*



Note: Standardized absolute mean differences of a selected subset of other covariates across profiles with and without smile*. Propensity score used for reweighing obtained using GBM model trained on all covariates in* Kiva data*.*

Figure 22: Propensity scores for *smile*



Note: Estimates of the propensity for an image to be a smile *using a GBM-based prediction with the full set of controls from* Kiva data.

Table 9: AIPW estimates of the impact of *style* features using alternative outcome and propensity models

name	ATE (reg.forest, B)	SE	ATE (GBM)	SE	ATE (GBM, B)	SE
Smile	8.81	0.83	3.62	0.56	2.21	0.96
Bodyshot	-8.39	0.75	-4.48	0.69	-4.14	0.76

Note: All regressions use all covariates from Kiva data. *Columns two* ATE (reg. forest, B) *and three show results based on the regression forests model for both the outcome and propensity, with observations reweighed using Li et al. (2018) method. Columns four* ATE (GBM) *and five are based on GBM predictions of the outcome and propensity model, and columns six* ATE (GBM, B) *and seven are also based on GBM but with additionally reweighing as in Li et al. (2018).*

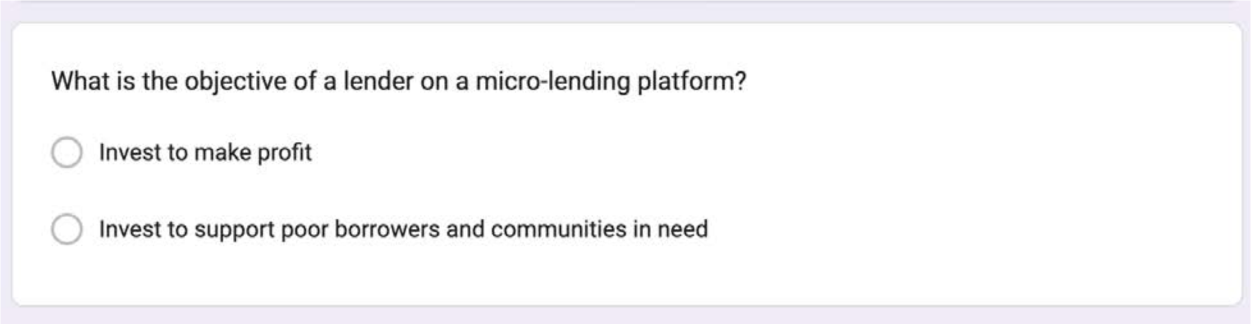
G Attention checks in the experiment

To check the quality of experimental data, we included attention checks in the survey. Attention checks are questions designed explicitly to detect inattentive responses through direct queries of attention or through questions designed to catch inattentive respondents (Abbey and Meloy (2017)). There are three purposes of the attention checks in our experimental setting First, attention checks ensure that the recruited subjects are fully informed of their roles in the correct context before subjects make decisions. Second, attention checks prevent the subjects from careless decision-making and help the recruited subjects make rational decisions. In addition, attention checks also give us the flexibility to filter the data in order to have high-quality ones, depending on whether we would like to tighten or loosen our criteria.

In order to avoid the attention checks themselves inducing a deliberative mindset and becoming a threat to the validity, we try to ask the subjects to recall detail in a previous image after they make the choice and the correct answer to that gives us the reason to believe that people have been paying rational attention to their choices.[29]

The Attention check 23 asks *What is the objective of a lender on a micro-lending platform?*. This question clarifies the lenders' role by differentiating the role between profit-making investors and non-profit investors. By answering this question correctly, the recruited subject understands that, as a donor in a non-profit micro-lending platform dedicated to expanding equal and reachable loan access, their goal should be supporting the poor borrowers and communities in need, instead of investing for profit (a prompt with this information was provided earlier in the survey).

Figure 23: Attention check 1



What is the objective of a lender on a micro-lending platform?

○ Invest to make profit

○ Invest to support poor borrowers and communities in need

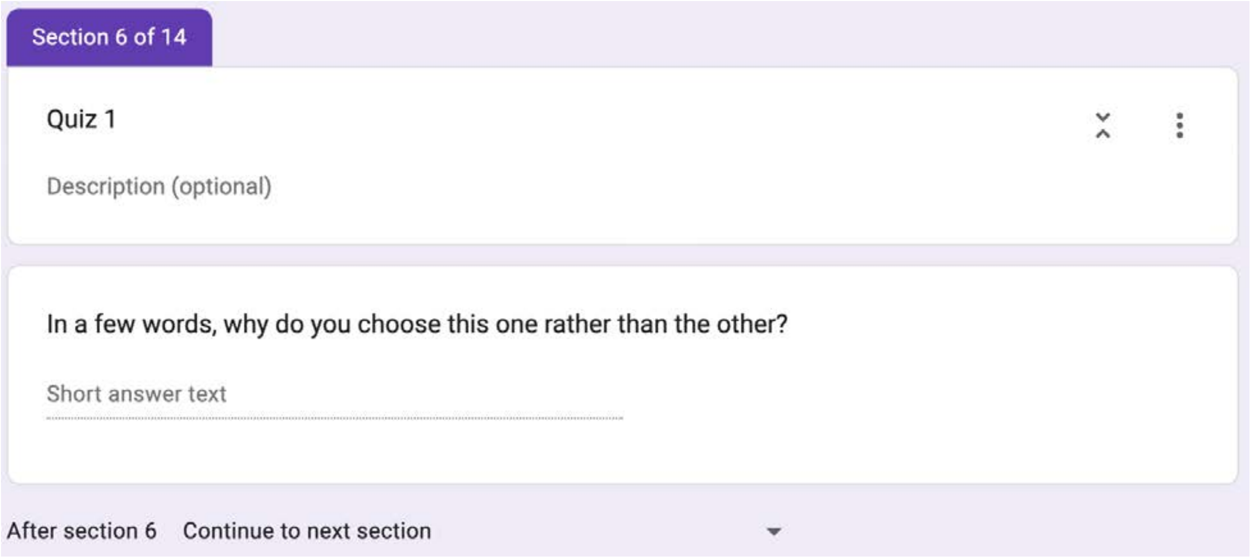
Attention check 24 and 25 are conducted in the format of a quiz. Attention check 24 is an open-ended query asking the subject for the reason of their decisions.[30] The last check is a multiple choice query asking about the occupation of the borrower on the previous slide.

Figures 26 and 27 show shares of subjects that responded correctly to Attention check 1 and 3. In both cases, correct response rates are above 90%. We take this as an indication that subjects were generally paying attention to their choices.

[29] Kung et al. (2018) encourage researchers to justify the use of attention checks without compromising scale validity

[30] Abbey and Meloy (2017) uses this type of attention checks and manipulation validations to detect inattentive respondents in primary empirical data collection

Figure 24: Attention check 2



Section 6 of 14

Quiz 1

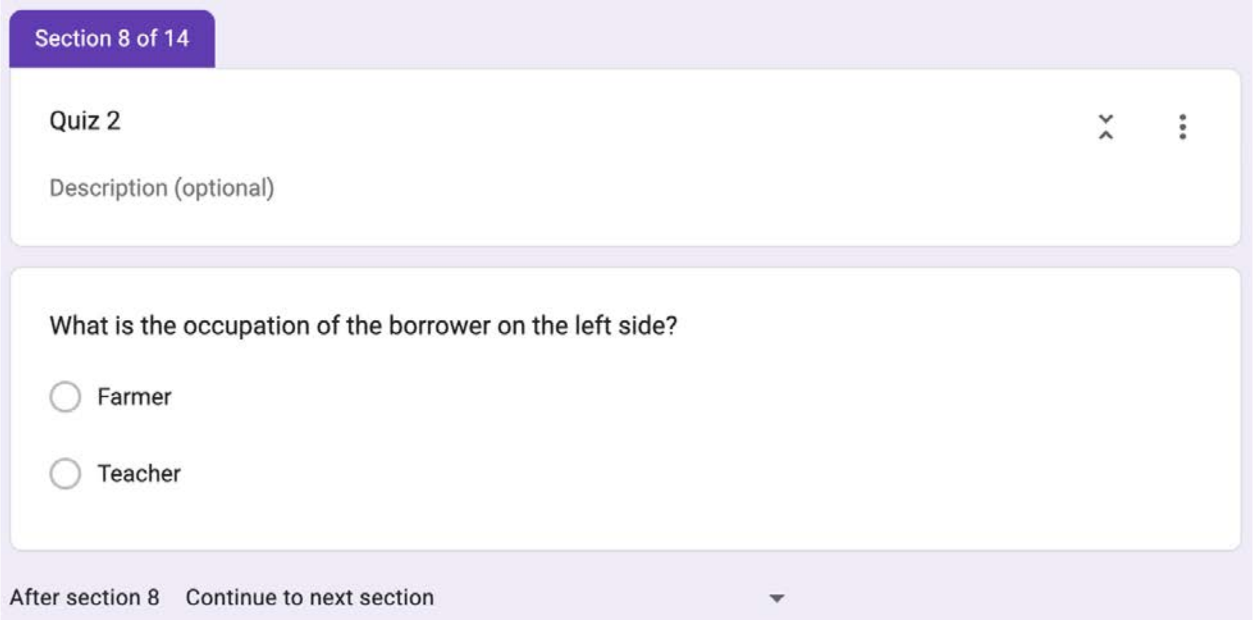
Description (optional)

In a few words, why do you choose this one rather than the other?

Short answer text

After section 6 Continue to next section

Figure 25: Attention check 3



Section 8 of 14

Quiz 2

Description (optional)

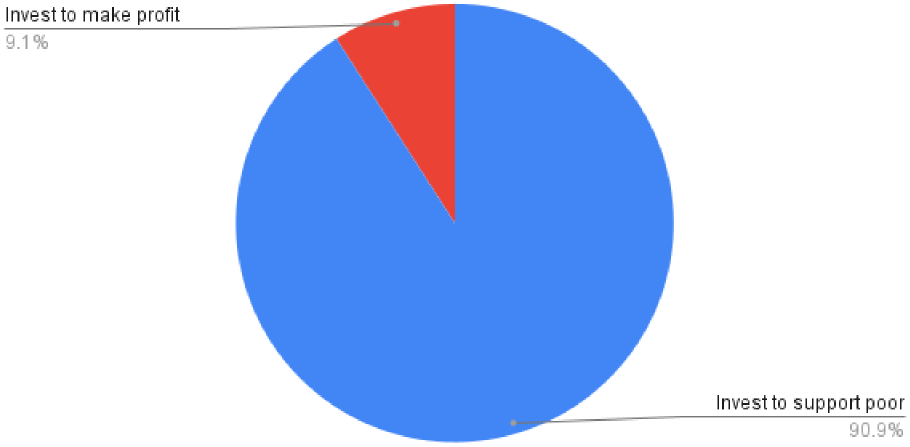
What is the occupation of the borrower on the left side?

○ Farmer

○ Teacher

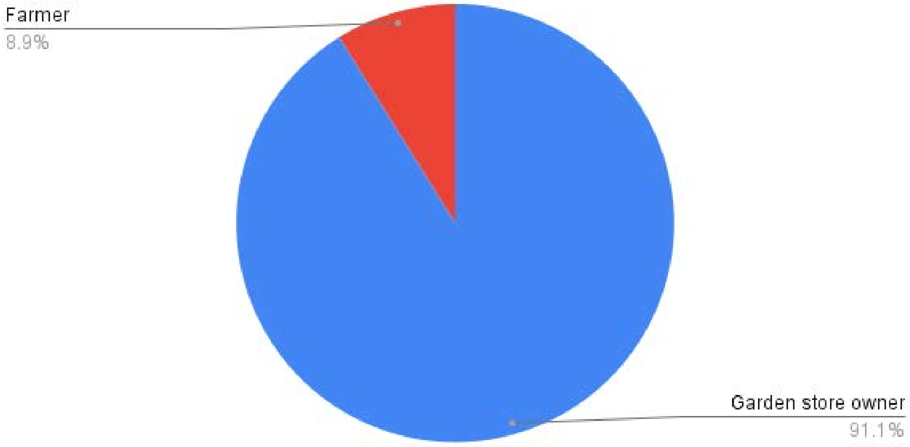
After section 8 Continue to next section

Figure 26: The proportion of correct answers (blue) to the object of a lender



Note: Count of What is the objective of a lender on a micro-lending platform

Figure 27: The proportion of correct answers (blue) to the borrower's occupation is shown on the previous page

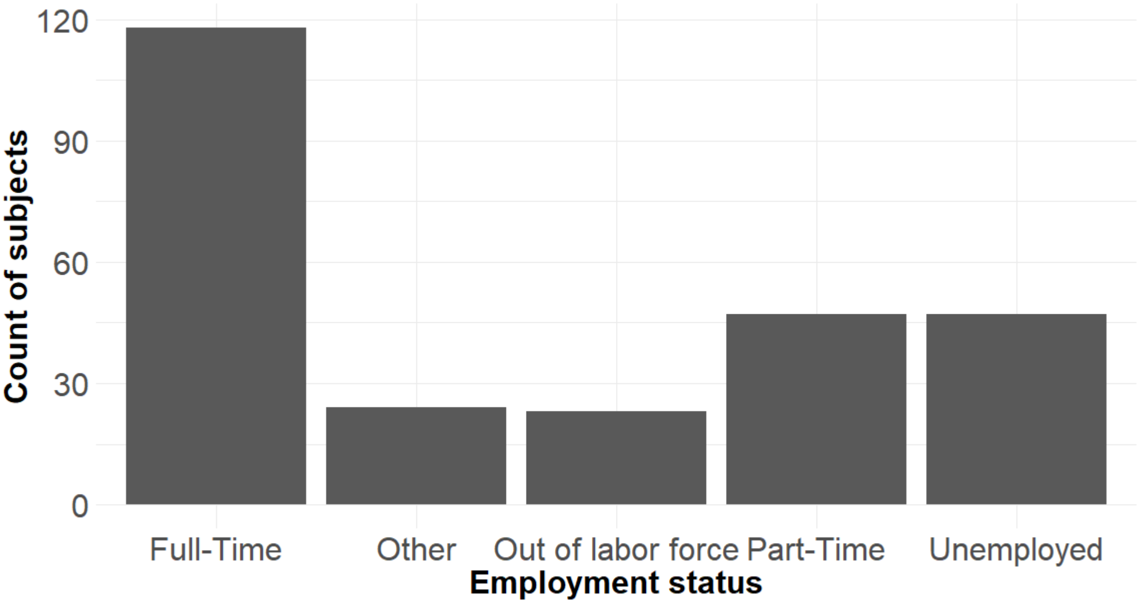


Note: Count of borrower's occupation shown in the previous page

H Summary statistics from the experiment

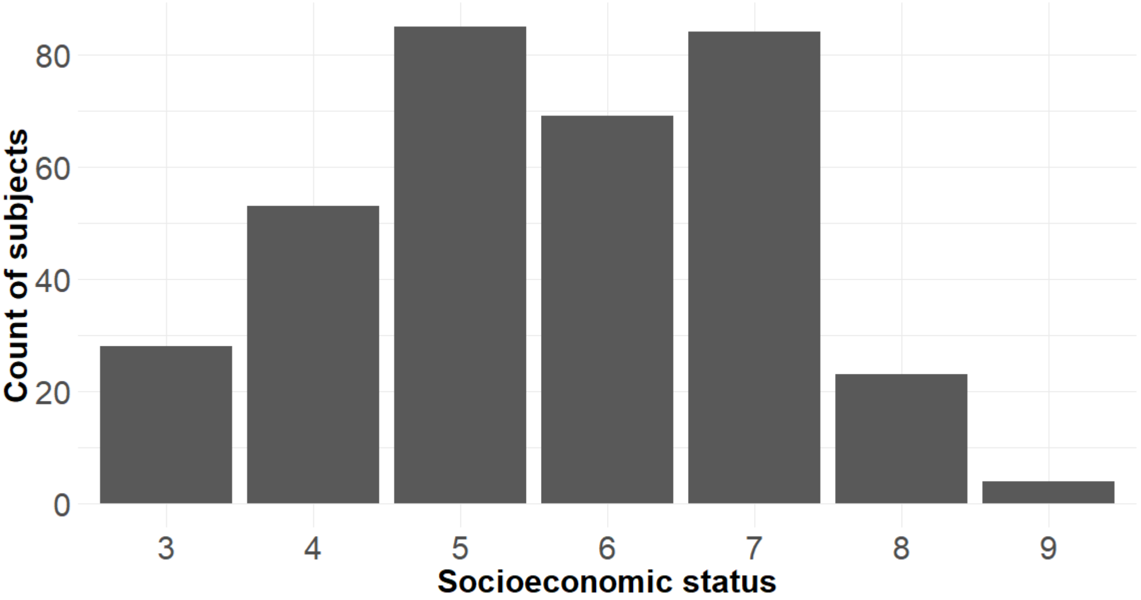
Figure 28 presents the employment status as reported by the subjects, there is a lot of missing data, to a large extent, this is due to the employment information being expired. Figure 29 shows the self-reported socioeconomic status, we screened participants to be of at least status 3.

Figure 28: Current employment status.



Note: Self-reported employment status. We drop observations where the data is unavailable (29%) and we group together full-time and 'starting new job soon' responses.

Figure 29: Self-assessed socio-economic status.



Note: Self-assessed socio-economic status. We required subjects to have at least a score of 3.

I Algorithm for counterfactual simulations

In this section, we describe the algorithm for generating outcomes under counterfactual policies in more detail. We divide the algorithm into two parts: (i) simulation of a market, and (ii) simulation of lenders' choices.

Algorithm 1 Simulation of a market

```
  $\tilde{\eta} \leftarrow U(\mathcal{N};22)$          ▷ Draw 22 fixed effects uniformly from the set of estimated fixed effects
  $\widetilde{male} \leftarrow \mathbf{E}_G[male|D(\tilde{\eta});22]$          ▷ Draw 22 gender realizations
  $\widetilde{bodyshot} \leftarrow \mathbf{E}_G[bodyshot|D(\tilde{\eta}),\widetilde{male};22]$
  $\widetilde{smile} \leftarrow \mathbf{E}_G[smile|D(\tilde{\eta}),\widetilde{male};22]$
  if $H \in \{Partialcompliance\}$ then
      if $\widetilde{bodyshot} == 1$ then
          $\widetilde{bodyshot} = B_{0.25}$          ▷ Bernoulli trial with $p = 0.25$
      end if
      if $\widetilde{smile} == 0$ then
          $\widetilde{smile} = B_{0.75}$
      end if
  end if
  $x \leftarrow (\tilde{\eta},\widetilde{male},\widetilde{bodyshot},\widetilde{smile})$

  if $H \in \{RestrictCompetition\}$ then
      $\mathcal{M} \leftarrow h(x;5)$
  else    $\mathcal{M} \leftarrow h(x;11)$     ▷ Draw borrowers from the pool following the probability function $h$
end if
$\mathcal{M} \leftarrow (\mathcal{M},\omega)$          ▷ add outside option
  return $\mathcal{M}$
```

Algorithm 1 proceeds in two steps, first, simulates the pool of borrowers and, second, samples from the pool to construct the market. Policies impact the distribution of the features in the pool (*partial compliance*), the size of the market (*Restrict competition*), and the probability of being sampled into the market (through the function h).

Once a market is simulated we determined lenders' choices with Algorithm 2. We first simulate the preferences of a lender, then compute the utility associates from different borrowers, and, finally, determined which borrower is selected.

Algorithm 2 Simulation of a lender choice

```
  $(\tilde{\alpha},\tilde{\beta},\tilde{\gamma}) \leftarrow (N(\alpha,sd_\alpha),N(\beta,sd_\beta),N(\gamma,sd_\gamma))$          ▷ draw preference parameters
  $\tilde{\epsilon} \leftarrow GEV$          ▷ draw random utility parameters for each borrowing campaign
  $u \leftarrow U(\mathcal{M};\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\epsilon})$          ▷ compute utilities from choosing any of the borrowers
  $choice \leftarrow max(u)$
  return $choice$
```