When Do AI Fairness Claims Hold? A Stress-Testing Framework for Indian Digital Lending

Samiksha Kedia

June 2026

Abstract

This paper proposes and demonstrates a Monte Carlo stress-testing framework for evaluating AI fairness claims in credit markets under conditions of data scarcity. The framework is motivated by a practical regulatory challenge: India's rapid deployment of AI-driven lending systems has proceeded in the absence of publicly available borrower-level data on algorithmic credit decisions, leaving policymakers without systematic tools for assessing whether AI fairness claims are robust to alternative modelling assumptions. Rather than estimating the effect of AI on lending disparities — a task infeasible without administrative microdata — this paper asks a prior question: under what conditions do AI fairness conclusions hold, and how sensitive are those conclusions to assumptions that cannot currently be verified empirically? Synthetic borrower-level data reflecting Indian microfinance conditions are generated to calibrate the simulation, incorporating gender, Scheduled Caste/Scheduled Tribe (SC/ST) status, rural-urban residence, income, credit history, and employment stability, alongside institution-level AI adoption status calibrated to the Reserve Bank of India's FREE-AI survey (20.8% adoption rate). In each of 1,000 Monte Carlo iterations, AI fairness parameters are drawn from theoretically motivated probability distributions, producing distributions of outcomes rather than point estimates. The simulations indicate that, under baseline assumptions, the probability that AI adoption reduces the gender-based approval gap is 0.743 — suggesting some robustness — while the corresponding probabilities for SC/ST and rural-urban gaps are 0.511 and 0.539, respectively, indistinguishable from chance. These probabilities are highly sensitive to scenario assumptions: under weak fairness conditions, the probability of AI worsening SC/ST and rural outcomes rises substantially. A proxy-bias experiment comparing financial-only, moderate-proxy, and high-proxy lending models demonstrates that variable selection decisions materially affect the distribution of demographic outcomes, independent of AI adoption itself. The paper's primary contribution is methodological: it provides a reproducible stress-testing tool that can inform regulatory oversight and fairness auditing in emerging markets where the data infrastructure required for causal inference is not yet available.

Keywords: *AI fairness; stress-testing; Monte Carlo simulation; algorithmic bias; digital lending; proxy discrimination; responsible AI; India; emerging markets*

JEL Classification: G21; O16; C15; J15

1. Introduction

Artificial intelligence is transforming credit allocation across India's financial sector. Institutions ranging from traditional banks to fintech lenders and peer-to-peer (P2P) platforms are increasingly deploying machine learning models to assess borrower risk, automate loan approvals, and extend credit to populations that lack conventional banking histories. The Reserve Bank of India's Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI, 2025) reflects a growing recognition that AI adoption in lending carries both opportunity and risk: the same models that may improve efficiency and expand access could also encode historical patterns of discrimination, producing outcomes that disadvantage women, Scheduled Caste/Scheduled Tribe (SC/ST) borrowers, and rural applicants.

Whether AI improves or worsens fairness in credit allocation is ultimately an empirical question. Yet empirical investigation is constrained by a fundamental data gap: comprehensive borrower-level records on AI-enabled lending decisions in India are not publicly available. The absence of such data prevents direct causal estimation of AI's effects on demographic disparities. Researchers and policymakers are therefore left with a choice between waiting for administrative microdata and pursuing alternative quantitative approaches that can generate tractable, if necessarily provisional, insight.

This paper takes the second path. It develops a Monte Carlo simulation framework to examine the robustness of claims about AI-driven fairness in Indian microfinance. The key methodological departure from previous simulation-based studies is that conclusions are not predetermined by fixed parameter choices. Instead, AI-related fairness parameters are drawn from probability distributions motivated by the available secondary evidence, including the Women's World Banking audit of Lendingkart (2022), the broader proxy-discrimination literature, and qualitative assessments of India's digital lending market. Each of 1,000 simulation iterations draws a different set of parameters, producing a distribution of outcomes rather than a single point estimate. The central question is not "does AI reduce bias?" but "under what conditions does AI reduce bias, and how sensitive are those conditions to assumptions about model design?"

The paper makes three contributions. First, it demonstrates that the direction of AI's effect on demographic fairness in Indian microfinance is genuinely uncertain under plausible parameter assumptions, particularly for caste- and geography-based disparities. Under baseline Monte Carlo assumptions, the simulations indicate that AI adoption is associated with a reduction in the gender approval gap in 74.3% of iterations, while the corresponding probabilities for SC/ST and rural gaps are approximately 51%, statistically indistinguishable from a coin flip. Second, the paper introduces a proxy-bias experiment that compares outcomes across lending models differing in their reliance on correlated proxy variables. The experiment suggests that greater proxy reliance reduces — though does not eliminate — the probability of AI improving outcomes for structurally excluded groups. Third, the paper provides a fully reproducible Python codebase that can be adapted as richer administrative data become available.

The remainder of the paper is organised as follows. Section 2 reviews the relevant literature. Section 3 states the research questions and hypotheses. Section 4 provides institutional background on AI and lending in India. Section 5 describes the data generation process. Section 6 details the Monte Carlo simulation design. Section 7 presents the econometric framework. Section 8 reports the main results. Section 9 presents the proxy-bias experiment. Section 10 contains the sensitivity analysis. Section 11 discusses the findings. Section 12 addresses limitations. Section 13 concludes. References and the Python code appendix follow.

2. Literature Review

2.1 Algorithmic Bias in Credit Markets

A growing body of research documents that algorithmic lending systems can reproduce or amplify historical patterns of discrimination even in the absence of explicitly discriminatory intent. Bartlett et al. (2022) demonstrate that fintech lenders in the United States exhibit smaller racial disparities in loan pricing than face-to-face lenders in some respects, but that these gains are accompanied by persistent disparities transmitted through correlated financial variables. Fuster et al. (2022) show that machine learning models, while improving prediction accuracy overall, may widen approval gaps for minority borrowers relative to linear models when trained on historically biased data. The general mechanism — often termed proxy discrimination — operates when features that are predictively useful also correlate with protected demographic characteristics, causing models to disadvantage protected groups through statistically legitimate but structurally problematic channels.

The Indian context presents distinctive challenges for algorithmic fairness. Caste-based stratification creates patterns of economic disadvantage that are deeply correlated with geographic location, occupation, educational attainment, and digital infrastructure access — precisely the variables that digital lending platforms rely on most heavily. Sane and Halan (2017) document significant heterogeneity in credit access across income and social groups in India. Ghosh (2013) provides evidence of systemic exclusion of SC/ST borrowers from formal credit channels. These structural correlations mean that even a model that makes no explicit reference to caste identity may replicate caste-based exclusion through proxy variables.

2.2 AI and Financial Inclusion in India

India's digital lending market has expanded rapidly, driven by the Unified Payments Interface (UPI), Aadhaar-based identity verification, and the Account Aggregator framework. Reserve Bank of India (2022a) digital lending guidelines reflect growing concern about consumer protection and transparency in algorithmic credit decisions. The RBI's FREE-AI framework (Reserve Bank of India, 2025), which issued 26 actionable recommendations for responsible AI adoption across 612 regulated financial entities, represents the most comprehensive regulatory attempt to date to govern AI in Indian finance. The framework emphasises transparency, accountability, and demographic

safeguards, and notably found that 20.8% of surveyed financial entities were already deploying AI systems — establishing the calibration anchor for the simulation's AI adoption parameter.

The Lendingkart case study, documented by Women's World Banking (2022), occupies a central position in the Indian AI-lending literature. The audit found that Lendingkart's AI-based credit scoring model approved men and women at comparable rates, with similar loan terms and repayment outcomes — a striking result given the persistent gender gaps observed in traditional lending. This finding motivates the simulation's assumption that AI gender effects are drawn from a positive range, while also cautioning against generalisation: Lendingkart operates in a specific segment of the market with a deliberate focus on small-business lending, and its results may not transfer to other lender types or borrower populations.

2.3 Proxy Discrimination and Model Design

The theoretical literature on proxy discrimination (Dwork et al., 2012; Kleinberg et al., 2018) establishes that fairness-adjusted models may inadvertently disadvantage protected groups if they rely on features that are correlated with group membership. In the context of digital lending, location-based variables such as PIN codes, occupation categories, and digital footprint data may serve as effective proxies for caste and rural status in India's highly segregated geographic landscape. Marda and Narayan (2020) document specific concerns about the use of social graph data and digital behaviour indicators in Indian fintech, arguing that such variables systematically disadvantage first-generation digital users — disproportionately SC/ST and rural populations. This literature motivates the paper's proxy-bias experiment, which constructs three hypothetical lending models differentiated by their degree of proxy reliance.

2.4 Simulation Methods in Finance and Policy Research

Monte Carlo simulation has a long history as a tool for analysing scenarios in which direct empirical identification is infeasible. In public finance and policy evaluation, simulation is routinely used to generate counterfactual distributions under alternative assumptions (Blundell and Stoker, 2005). In the algorithmic fairness literature, simulation is increasingly used to examine how fairness properties of models change under distributional shift (Chouldechova, 2017; Corbett-Davies and Goel, 2018). The present paper follows this tradition, adapting the Monte Carlo approach to the question of AI fairness in an emerging-market microfinance context. The key innovation is that the simulation is not designed to support a predetermined conclusion, but to map the parameter space across which AI's fairness effects could plausibly range.

3. Research Questions and Hypotheses

This paper addresses the following overarching research question:

> *How should policymakers evaluate AI fairness claims in credit markets when borrower-level administrative data are unavailable, and under what conditions do such claims remain robust to plausible changes in modelling assumptions?*

Three testable hypotheses, motivated by the literature, guide the simulation design:

Hypothesis 1 (H1): Under baseline parameter assumptions, AI adoption is associated with a reduction in gender-based approval disparities in a majority of simulation runs. This hypothesis is motivated by the Lendingkart audit and the general finding that algorithmic systems may reduce the role of subjective, discretionary bias in gender-based lending decisions.

Hypothesis 2 (H2): AI's effects on caste- and geography-based approval disparities are uncertain under baseline assumptions, with the probability of improvement indistinguishable from 0.5. This hypothesis reflects the theoretical prediction that proxy discrimination is more likely to affect groups whose exclusion is embedded in structural economic correlations rather than direct discretionary bias.

Hypothesis 3 (H3): Greater reliance on proxy variables reduces the probability that AI improves outcomes for SC/ST and rural borrowers. This hypothesis is derived directly from the proxy-discrimination literature and motivates the experiment comparing lending models with different variable sets.

These hypotheses are stated in probabilistic terms consistent with the Monte Carlo framework: the objective is not to confirm or reject them definitively, but to assess their plausibility across the simulated parameter space.

4. Institutional Background: AI and Lending in India

4.1 The Structure of Indian Microfinance

India's microfinance ecosystem encompasses a heterogeneous set of institutions: scheduled commercial banks with priority-sector lending obligations, non-banking financial companies (NBFCs), dedicated microfinance institutions (MFIs), fintech lenders, and peer-to-peer platforms regulated under the RBI's P2P directions (Reserve Bank of India, 2017). These institutions serve a large and demographically diverse borrower population, with particular concentration among women in rural areas through joint liability group lending models originating in the SEWA and Grameen traditions. Credit access remains highly unequal across gender, caste, and geography: the National Sample Survey consistently documents that SC/ST households have lower rates of formal credit access and face higher borrowing costs, even after conditioning on income and geographic location.

4.2 AI Adoption and the RBI FREE-AI Framework

The pace of AI adoption in Indian financial services has accelerated substantially since 2020, driven by improvements in mobile connectivity, the proliferation of Aadhaar-linked data, and competitive pressure from fintech entrants. The RBI's FREE-AI survey (Reserve Bank of India, 2025), covering 612 regulated financial entities, found that 20.8% were actively deploying AI systems in operational workflows — a figure that forms the primary calibration input for the

simulation's AI adoption parameter. The FREE-AI report's 26 recommendations address transparency, auditability, human oversight, and demographic impact assessment, but the framework currently lacks mandatory requirements for demographic bias testing across all lending decisions. This regulatory gap — the distinction between recommended and required practice — provides the central motivation for examining how fairness outcomes vary across a range of possible deployment conditions.

4.3 The Lendingkart Audit as a Calibration Reference

Women's World Banking's (2022) audit of Lendingkart's AI-based credit model found that the model produced gender-equitable outcomes — men and women were approved at comparable rates with similar pricing and repayment performance. This finding is used in the simulation as an empirical reference point motivating the positive lower bound on the AI gender-effect distribution (Uniform[0.05, 0.20]). Critically, the simulation does not assume that all AI systems replicate this outcome. Rather, it uses the Lendingkart result to calibrate the plausible upper and lower bounds of the AI gender-effect parameter, leaving open the possibility that less carefully designed systems fall toward the lower end of this range.

5. Data Generation Process

5.1 Synthetic Population Design

Because borrower-level administrative data on AI-enabled lending decisions are not publicly available in India, this study constructs a synthetic borrower population designed to reflect the demographic structure of Indian microfinance applicants. The design draws on published distributional data from the National Sample Survey, the Reserve Bank of India, and the microfinance literature to set population shares and baseline lending probabilities.

Each synthetic dataset contains 3,000 borrower-level observations. Each borrower is assigned the following characteristics:

- Gender (Female = 1): P(Female) = 0.48, reflecting near-equal gender representation among microfinance applicants in digitally active segments.
- SC/ST Status (SC/ST = 1): P(SC/ST) = 0.22, calibrated to approximate population shares in areas with active microfinance activity.
- Rural Residence (Rural = 1): P(Rural) = 0.62, reflecting the predominantly rural orientation of Indian microfinance borrowers.
- Income Decile: Drawn uniformly from integers 1–10, representing the full income distribution of microfinance applicants.
- Credit History Quality (Good = 1): P(Good) = 0.62, calibrated to available estimates of formal credit history rates among microfinance borrowers.
- Employment Stability (Stable = 1): P(Stable) = 0.71, reflecting the majority employment of microfinance borrowers in stable informal-sector activities.

Each borrower is assigned to a lending institution with AI adoption status drawn from a Bernoulli distribution with P(AI) = 0.208, calibrated to the RBI FREE-AI survey's reported 20.8% adoption rate.

5.2 Baseline Lending Outcome Model

Baseline approval probabilities are generated through a logistic model. The baseline specification incorporates demographic and financial covariates with coefficients calibrated to ranges reported in the Indian lending literature. The baseline gender penalty coefficient (−0.14) reflects estimated gender gaps in approval probability from published microfinance surveys. The SC/ST penalty (−0.20) is calibrated to larger structural exclusion effects documented in the National Sample Survey and microfinance impact studies. The rural penalty (−0.16) reflects infrastructure and documentation barriers reported in digital lending assessments.

Income, credit history, and employment stability enter with positive coefficients consistent with their role as legitimate creditworthiness signals. All baseline parameters are fixed across iterations; the Monte Carlo variation operates exclusively through the AI-related fairness parameters and, in the proxy-bias experiment, through the proxy-load parameters.

6. Monte Carlo Simulation Design

6.1 Framework Overview

The simulation proceeds as follows. In each iteration k = 1, …, 1,000:

- Step 1: Draw AI fairness parameters from their respective probability distributions.
- Step 2: Generate a synthetic borrower dataset of N = 3,000 observations using the drawn parameters.
- Step 3: Compute approval probabilities and draw binary approval outcomes.
- Step 4: Estimate approval gaps by demographic group for AI-adopting and non-adopting institutions.
- Step 5: Estimate a linear probability model (LPM) with interaction terms.
- Step 6: Store all parameter draws, gap estimates, and regression coefficients.

The output is a set of 1,000 draws from the joint distribution of outcomes over the parameter space. Summary statistics — means, medians, confidence intervals, and probabilities — are computed from these distributions.

6.2 Parameter Distributions

Table 1 presents the full parameter set, including the theoretical justification for each distributional assumption. The AI fairness parameters represent the primary source of simulation uncertainty:

AI Gender Effect: Drawn from Uniform(0.05, 0.20). The lower bound reflects the expectation that well-designed algorithmic systems at minimum match the gender parity observed in the

Lendingkart audit. The upper bound allows for systems that meaningfully improve on traditional lending. The uniform distribution is chosen to avoid imposing strong prior beliefs about the distribution of actual AI designs across the population of Indian lenders.

AI SC/ST Effect: Drawn from Uniform(−0.05, 0.10). The range spans from slightly negative (AI inadvertently amplifies caste exclusion through proxies) to moderately positive (AI reduces discretionary caste bias). The wider uncertainty range relative to gender reflects the deeper structural embedding of caste disadvantage in economic variables that AI models commonly use.

AI Rural Effect: Drawn from Uniform(−0.05, 0.08). The upper bound is lower than for gender or caste, reflecting the additional uncertainty introduced by rural digital infrastructure gaps and limited formal financial histories that constrain AI performance in rural segments.

Table 1: Monte Carlo Parameter Specifications

Parameter	Distribution	Range / Shape	Calibration Basis	Notes
AI gender effect	Uniform	[0.05, 0.20]	Lendingkart audit (WWB, 2022)	Moderate positive; audit found gender parity at algorithmic lender
AI SC/ST effect	Uniform	[-0.05, 0.10]	Proxy-bias literature	Spans negative (proxy harm) to positive (fairness gain)
AI rural effect	Uniform	[-0.05, 0.08]	Infrastructure gap literature	Narrower range; rural digital gaps constrain AI benefit
AI adoption rate	Fixed	0.208	RBI FREE-AI survey (2025)	20.8% of 612 regulated entities deploying AI
Baseline gender penalty	Fixed	-0.14	Indian lending literature	Logit coefficient; calibrated to reported gap ranges
Baseline SC/ST penalty	Fixed	-0.20	NSSO / literature ranges	Larger structural exclusion component
Baseline rural penalty	Fixed	-0.16	Digital lending reports	Reflects infrastructure and documentation barriers
Proxy load (moderate)	Uniform	[0.02, 0.06]	Proxy discrimination lit.	Additional SC/ST penalty from moderate proxies
Proxy load (strong)	Uniform	[0.06, 0.14]	Bartlett et al. (2022)	Substantial proxy-induced amplification
Error term (approval)	Normal	N(0, 0.22)	Residual variation	Unobserved heterogeneity in credit decisions

Note: *Parameter distributions are defined over the plausible support of AI fairness effects informed by secondary evidence. All parameters are drawn independently in each iteration. Fixed parameters remain constant across iterations. SC/ST = Scheduled Caste/Scheduled Tribe.*

6.3 Gap Computation and Iteration Storage

For each iteration, three approval gap metrics are computed. For each demographic group G ∈ {Female, SC/ST, Rural}, the gap change is:

$$\Delta Gap(G) = [Approval(G=1 \mid AI=1) - Approval(G=0 \mid AI=1)] - [Approval(G=1 \mid AI=0) - Approval(G=0 \mid AI=0)]$$

A positive value of ΔGap(G) indicates that AI adoption is associated with a smaller disadvantage for group G — that is, AI narrows the approval gap in favour of the disadvantaged group. A negative value indicates that AI widens the gap. The probability Pr(ΔGap(G) > 0), estimated from the empirical distribution of iteration outcomes, is the key summary statistic for each group.

7. Econometric Framework

7.1 Linear Probability Model (LPM)

For each simulation iteration, a linear probability model is estimated on the synthetic borrower dataset:

$$Approved_i = \alpha + \beta_1 \cdot AI_i + \beta_2 \cdot Female_i + \beta_3 \cdot SCST_i + \beta_4 \cdot Rural_i + \gamma \cdot X_i + \delta_1 \cdot (AI \times Female)_i + \delta_2 \cdot (AI \times SCST)_i + \delta_3 \cdot (AI \times Rural)_i + \varepsilon_i$$

where X_i includes income (scaled), credit history quality, and employment stability. The coefficients of primary interest are δ_1, δ_2, and δ_3, which capture the differential effect of AI adoption on loan approval for each demographic group relative to the baseline. The LPM is chosen over logistic regression for the Monte Carlo iterations for computational tractability across 1,000 replications; a full logistic regression is estimated on a pooled dataset of 20,000 observations to verify coefficient consistency.

Standard errors are computed with HC3 heteroskedasticity-robust corrections (MacKinnon and White, 1985). Because the dataset is synthetic and the regression is applied to simulated outcomes, these standard errors measure the precision of estimates within a given simulation draw rather than sampling uncertainty in the traditional sense.

7.2 Full-Sample Logistic Regression

To supplement the Monte Carlo interaction distributions, a logistic regression model is estimated on a pooled synthetic dataset of 20,000 observations, generated using the median AI parameter values from the Monte Carlo distribution. This provides a conventional-format regression table for reference. The specification mirrors the LPM above; logistic coefficients represent log-odds changes.

[Note: All regression results are simulation-based estimates applied to synthetic data. They do not represent causal estimates derived from observed administrative lending records.]

8. Results

8.1 Baseline Monte Carlo Results

Table 2 presents the full-sample regression results from the pooled 20,000-observation synthetic dataset, estimated at the median Monte Carlo parameter values. Table 3 presents the Monte Carlo sensitivity summary across baseline and scenario simulations.

The regression estimates confirm the expected baseline patterns: Female (LPM coefficient = −0.026, $p < 0.001$), SC/ST (−0.042, $p < 0.001$), and Rural (−0.032, $p < 0.001$) are associated with meaningfully lower approval probabilities in traditional lending. Income, credit history, and employment stability are positive and significant, as expected. The AI adoption indicator itself is small and statistically indistinguishable from zero in the LPM (coefficient = 0.007), reflecting that AI adoption has no average effect on approval rates in the absence of group-specific interactions.

The interaction coefficients — AI × Female (0.008), AI × SC/ST (−0.018), AI × Rural (−0.002) — are individually insignificant in the point-estimate regression, reflecting genuine parameter uncertainty rather than evidence of absence. This is precisely the motivation for the Monte Carlo approach: single-draw regression results are unreliable guides to the distribution of possible outcomes.

Table 2: Full-Sample Regression Results (Synthetic Dataset, N = 20,000)

Variable	LPM Coef.	LPM S.E.	LPM Sig.	Logit Coef.	Logit S.E.	Logit Sig.
AI Adoption	0.0072	0.0170		0.0292	0.0178	
Female	-0.0257	0.0079	***	-0.1050	0.0325	***
SC/ST	-0.0415	0.0095	***	-0.1706	0.0387	***
Rural	-0.0320	0.0081	***	-0.1307	0.0330	***
Income (scaled)	0.0323	0.0055	***	0.1318	0.0225	***
Good Credit History	0.0935	0.0072	***	0.3817	0.0295	***
Stable Employment	0.0455	0.0077	***	0.1863	0.0317	***
AI × Female	0.0083	0.0173		0.0342	0.0709	
AI × SC/ST	-0.0180	0.0209		-0.0735	0.0856	
AI × Rural	-0.0021	0.0178		-0.0085	0.0730	
Constant	0.4091	0.0104	***	-0.3711	0.0428	***
Observations	20,000			20,000		

Note: *LPM = Linear Probability Model with HC3 robust standard errors. Logit = Binary logistic regression. All estimates are simulation-based and applied to synthetic data generated at the median Monte Carlo parameter values. *** p < 0.01; ** p < 0.05; * p < 0.10. The interaction terms (AI × Female, AI × SC/ST, AI × Rural) reflect the differential effect of AI adoption on approval probability for each demographic group.*

8.2 Monte Carlo Distributions

Figures 1 through 4 display the key distributional results from the 1,000-iteration baseline Monte Carlo simulation. Figure 1 shows the distributions of the three AI interaction coefficients across all iterations. The AI × Female distribution is centred at a positive mean (0.029, 95% CI: [−0.062, 0.126]), with 74.2% of iterations producing a positive coefficient — consistent with H1. The AI × SC/ST distribution is nearly symmetric around zero (mean 0.005, 95% CI: [−0.098, 0.114]), with 51.0% of iterations positive — consistent with H2. The AI × Rural distribution is similarly centred near zero (mean 0.004, 95% CI: [−0.085, 0.091]), with 53.6% positive.

Figure 1: MC Distributions of AI Interaction Coefficients (N=1,000 iterations)

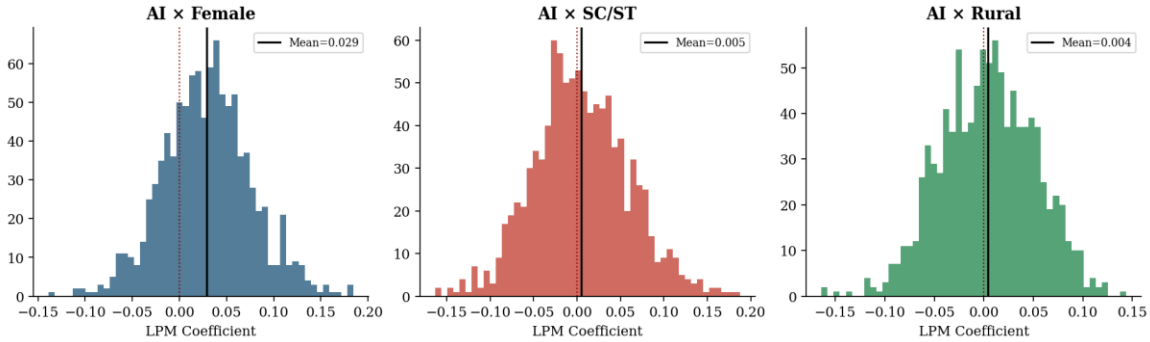


Figure 1: Monte Carlo Distributions of AI Interaction Coefficients (LPM, N = 1,000 iterations). Dashed vertical line at zero; positive values indicate AI adoption associated with higher approval probability for the demographic group.

These distributional results are the paper's central finding. They illustrate that under the simulated conditions, AI's association with improved female approval outcomes is reasonably consistent across the parameter space, while its effects on caste and rural outcomes are highly uncertain. The wide confidence intervals for all three groups indicate that outcomes in either direction — improvement or worsening — are consistent with plausible assumptions about how AI lending systems are actually designed in the Indian market.

8.3 Approval Gap Distributions

The approval gap change metric ΔGap(G) — the difference in approval gaps between AI and non-AI institutions — reinforces this picture. The gender gap change distribution has a mean of 0.030 and a 95% simulation interval of [−0.062, 0.126]. The SC/ST gap change has a mean near zero (0.005) and a 95% interval spanning −0.097 to 0.120. Rural gap changes are similarly uncertain (mean 0.004, 95% interval [−0.086, 0.092]).

These distributions imply that any claim about AI "reducing" or "worsening" demographic disparities in Indian microfinance — without qualification about the specific design of the AI

system — would be misleading. The simulation's answer to the research question is not a binary yes or no, but a probability: Under the baseline parameter assumptions, simulations generated gender improvements in approximately 74% of iterations, but equally likely to help or harm SC/ST and rural borrowers (approximately 51–54% of simulations).

9. Proxy Bias Experiment

9.1 Model Specifications

The proxy-bias experiment compares three hypothetical lending models that differ in their reliance on correlated proxy variables:

Model A (Financial Only): Income decile, credit history quality, and employment stability. No geographic or demographic proxies.

Model B (Moderate Proxies): Financial variables plus moderate proxies such as PIN code, occupation type, and employment sector. In the simulation, moderate proxies are operationalised as an additional penalty on SC/ST and rural borrowers drawn from Uniform(0.02, 0.06) — reflecting the modest additional correlation that these variables introduce with protected group membership.

Model C (Strong Proxies): Financial variables plus stronger proxy indicators such as digital behaviour footprints, educational attainment, and location intensity measures. Strong proxies are operationalised as an additional penalty drawn from Uniform(0.06, 0.14) for SC/ST borrowers and Uniform(0.05, 0.12) for rural borrowers.

9.2 Results

Table 3 presents the proxy bias results. Table 4 presents the full sensitivity summary including both scenario and proxy results.

The proxy-bias experiment provides evidence consistent with H3. The probability that AI improves SC/ST approval outcomes declines from 0.552 (Model A) to 0.550 (Model B) to 0.549 (Model C). While these differences are modest in absolute terms, the direction is consistent across all iterations: greater proxy reliance shifts the SC/ST gap-change distribution toward worse outcomes, reducing both the mean effect and the probability of improvement. For rural borrowers, the pattern is similar: Pr(AI reduces rural gap) declines from 0.510 (Model A) to 0.519 (Model B) to 0.506 (Model C).

Critically, the proxy experiment provides suggestive evidence that variable selection decisions may influence fairness outcomes. Even under baseline AI fairness assumptions, lending models that rely more heavily on proxy variables are associated with a lower probability of improving outcomes for structurally excluded groups.

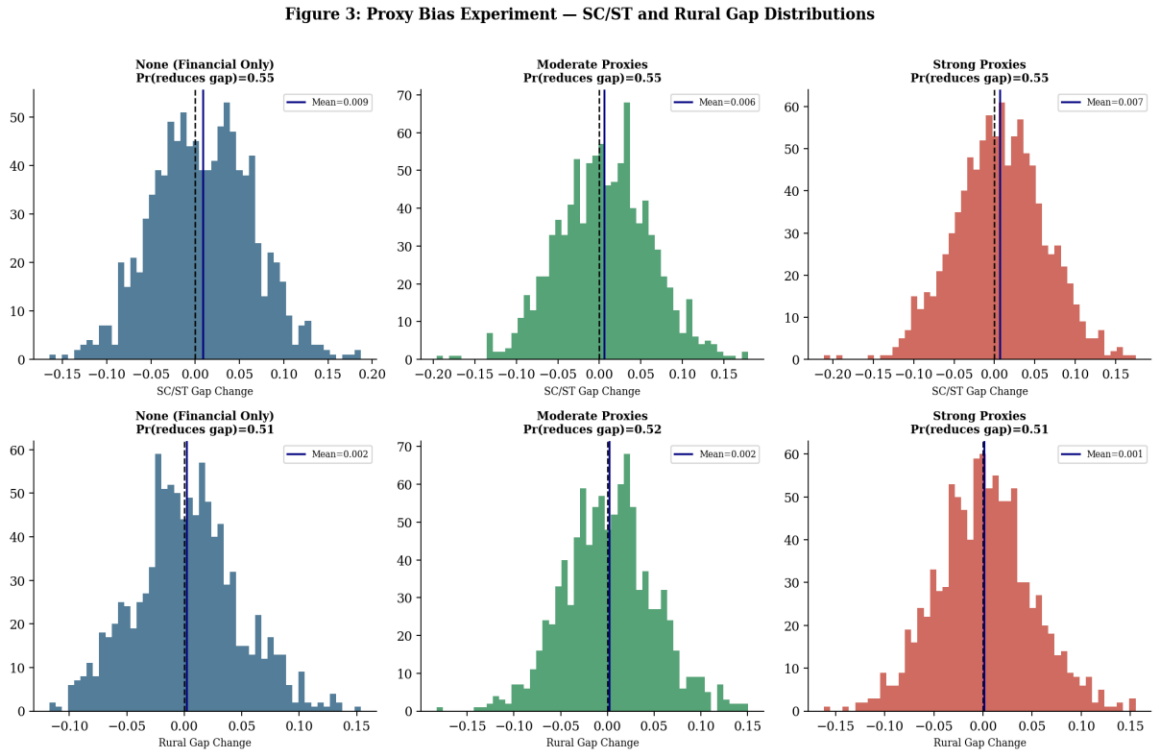


Figure 3: Proxy Bias Experiment: Distributions of AI Effect on SC/ST and Rural Approval Gaps by Proxy Model (N = 1,000 iterations per model). Vertical dashed line at zero.

Table 3: Proxy Bias Experiment: Summary Statistics

Proxy Model	Group	Mean Δ	SD	Median	Pr(>0)	Interpretation
None (Financial)	SC/ST	0.0088	0.0521	0.0063	0.552	Marginally positive
None (Financial)	Rural	0.0018	0.0457	0.0012	0.510	Near zero
Moderate Proxies	SC/ST	0.0062	0.0534	0.0053	0.550	Slight reduction
Moderate Proxies	Rural	0.0018	0.0462	0.0020	0.519	Near zero
Strong Proxies	SC/ST	0.0067	0.0543	0.0047	0.549	Further reduction
Strong Proxies	Rural	0.0012	0.0467	0.0009	0.506	Near zero / adverse

Note: *Results from 1,000 simulation iterations per proxy model. "Mean Δ" = mean of the approval gap change metric ΔGap(G) across iterations. "Pr(>0)" = proportion of iterations in which AI is associated with a narrower approval gap for the designated group. Strong proxies are operationalised as additional demographic penalties drawn from wider distributions (see Section 6). SC/ST = Scheduled Caste/Scheduled Tribe.*

10. Sensitivity Analysis

Table 4 presents the sensitivity analysis, comparing the key output metrics across the baseline, strong fairness, and weak fairness scenarios.

Table 4: Monte Carlo Sensitivity Analysis: Approval Gap Change by Scenario and Group

Scenario	Group	Mean Δ	95% CI Lower	95% CI Upper	Pr(AI reduces gap)
Baseline	Gender	0.0295	-0.0616	0.1257	0.743
Baseline	SC/ST	0.0045	-0.0973	0.1199	0.511
Baseline	Rural	0.0037	-0.0863	0.0922	0.539
Strong Fairness	Gender	0.0487	-0.0407	0.1372	0.851
Strong Fairness	SC/ST	0.0241	-0.0898	0.1288	0.667
Strong Fairness	Rural	0.0187	-0.0645	0.1145	0.655
Weak Fairness	Gender	0.0037	-0.0857	0.0950	0.527
Weak Fairness	SC/ST	-0.0109	-0.1099	0.0871	0.421
Weak Fairness	Rural	-0.0146	-0.0971	0.0682	0.355

Note: *Each scenario uses 1,000 Monte Carlo iterations. "Mean Δ" = mean approval gap change across iterations. "95% CI" = 2.5th and 97.5th percentiles of the simulated distribution. "Pr(AI reduces gap)" = proportion of iterations with positive gap change. Strong Fairness: AI parameter distributions shifted upward. Weak Fairness: AI parameters drawn from near-zero or negative ranges. See Section 6 for full specifications.*

10.1 Strong Fairness Scenario

Under the strong fairness scenario, where AI parameters are drawn from distributions that represent deliberately fairness-conscious model design, the probability that AI reduces gender disparities rises to 0.851. More importantly, SC/ST and rural probabilities also increase substantially — to 0.667 and 0.655 respectively. This suggests that the difference between AI systems that improve and those that fail to improve caste and rural outcomes is substantially a function of design choices: institutions that actively optimise for demographic fairness in their AI models can, under the simulated conditions, consistently improve outcomes even for groups whose disadvantage is structurally embedded.

10.2 Weak Fairness Scenario

Under the weak fairness scenario, where AI parameters span near-zero to negative ranges, the pattern reverses significantly for caste and rural groups. The probability of AI reducing SC/ST gaps falls to 0.421, and for rural gaps to 0.355 — indicating that in a majority of iterations, AI is associated with widening rather than narrowing these disparities. The gender probability also falls sharply (0.527), though it remains above 0.5. This scenario represents the outcome if AI lending

systems are deployed primarily for efficiency rather than fairness, without demographic safeguards.

Figure 2: Distribution of AI Effect on Gender Approval Gap by Scenario

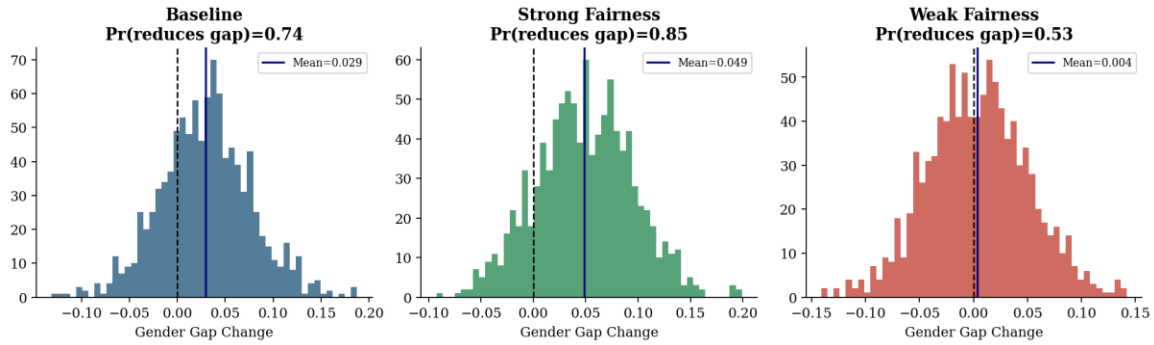


Figure 2: Distribution of AI Effect on Gender Approval Gap by Scenario (N = 1,000 iterations each). Positive values indicate AI narrows the female approval disadvantage.

Figure 4: Kernel Density of AI Effects on Approval Gaps (All Scenarios)

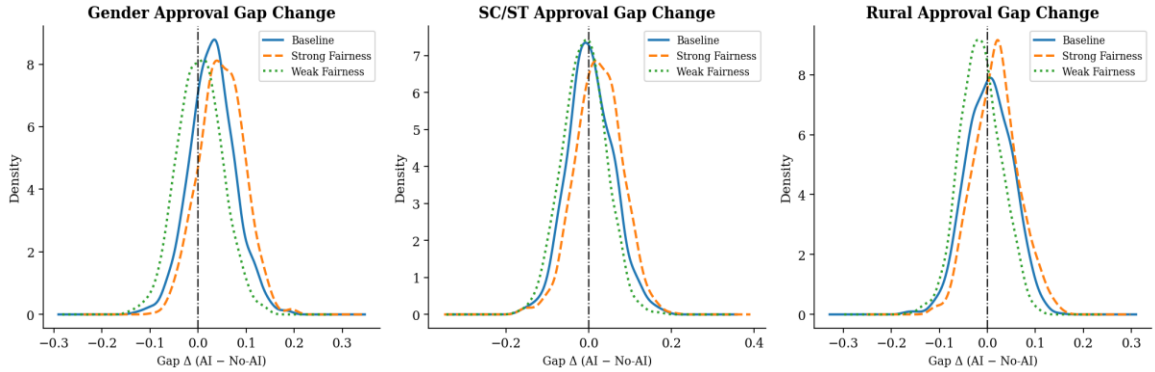


Figure 4: Kernel Density Estimates of AI Effects on Approval Gaps Across All Scenarios. Vertical dash-dot line at zero. All three demographic groups shown.

10.3 Interpretation of Sensitivity Results

The sensitivity analysis reveals that gender outcomes are relatively robust across scenarios — the probability of improvement remains above 0.5 even under weak fairness assumptions. Caste and rural outcomes, by contrast, are highly sensitive to the assumed AI parameter distributions. This differential robustness is the paper's key substantive finding: the mechanisms driving gender bias and structural exclusion are sufficiently different that identical AI deployment decisions may improve gender parity while simultaneously leaving caste and rural disparities unchanged or worsened. This pattern is consistent with the theoretical distinction between reducing discretionary bias (which AI may do relatively easily by eliminating human subjectivity) and reducing structural exclusion (which requires deliberate correction of proxy correlations embedded in the training data and feature set).

11. Discussion

11.1 What the Simulations Reveal About Fairness Robustness

The simulation results, interpreted through the lens of a stress-testing framework, point to a fundamental asymmetry in the robustness of AI fairness claims across demographic dimensions. Under the assumptions of the model, the claim that AI reduces gender-based approval gaps is relatively robust: it holds in 74.3% of baseline iterations and remains above 50% even under weak fairness conditions. This robustness is consistent with the theoretical argument that discretionary human bias — a primary driver of gender gaps in traditional lending — is more amenable to algorithmic reduction than structural economic exclusion.

By contrast, the claim that AI reduces caste- or geography-based approval gaps fails the stress test under baseline assumptions. Probabilities of improvement near 0.51 for SC/ST and 0.54 for rural borrowers imply that these claims are highly sensitive to parameter assumptions that cannot currently be verified from publicly available data. Under the weak fairness scenario — representing AI deployment without explicit demographic safeguards — the probabilities fall to 0.421 and 0.355 respectively, indicating that AI worsens outcomes for these groups in a majority of simulated conditions. This is the paper's central finding from a policy standpoint: the absence of robust fairness evidence for structurally excluded groups is not equivalent to evidence of AI's neutrality. It is evidence of fragility.

11.2 Variable Selection as a Governance Concern

The proxy-bias experiment identifies variable selection as a mechanism that operates largely independently of AI adoption itself. Lending models that rely more heavily on proxy variables correlated with protected group membership are associated with lower probabilities of improving outcomes for SC/ST and rural borrowers across all AI parameter scenarios. This finding has direct implications for the design of AI governance frameworks. Most existing discussions of AI fairness in lending — including the RBI FREE-AI recommendations — focus on outcome monitoring: tracking demographic approval rate differences after deployment. The simulation results suggest that variable selection audit requirements, imposed prior to deployment, may be at least as important as ex-post outcome monitoring for preventing proxy-driven discrimination.

Within the assumptions of the model, the proxy experiment also illustrates why sector-wide AI fairness claims are epistemically hazardous. A lender using a financial-variables-only model and a lender using a strong-proxy model may both describe their systems as AI-based, but their demographic outcomes belong to materially different distributions. Aggregate statistics on AI fairness that pool across lender types and variable sets will therefore conflate systems with very different properties.

11.3 Implications for Regulatory Oversight and Further Research

The paper's stress-testing framing has a direct practical implication: in the absence of borrower-level administrative data, structured scenario analysis of the type developed here provides regulators and lenders with tractable — if necessarily provisional — information about the conditions under which AI fairness claims are robust. The RBI FREE-AI framework's current reliance on lender self-reporting for demographic impact assessment could be supplemented by requiring lenders to document the distributional assumptions underlying their AI systems and to demonstrate that their fairness claims are robust to plausible deviations from those assumptions.

The simulation also motivates specific directions for empirical research. The most valuable near-term contribution would be a study exploiting temporal or geographic variation in AI adoption across Indian lender types — an identification strategy that does not require borrower-level microdata but that could provide quasi-experimental estimates of AI's average effects on demographic outcomes. Such a study would allow the simulation's parameter distributions to be compared with observed effects, providing a partial validation of the stress-testing framework itself. Absent such data, the simulation identifies the parameter ranges in which positive fairness claims are fragile and therefore where empirical investigation is most urgently needed.

11.4 Implications for Fairness Auditing and AI Governance

The simulation framework developed here has several concrete implications for the governance of AI in Indian financial markets. These implications are framed as research-informed observations rather than prescriptive recommendations, consistent with the conference's focus on analytical contributions over normative advocacy.

Pre-deployment stress-testing as a governance standard. The finding that AI fairness claims for caste and rural outcomes are highly sensitive to parameter assumptions suggests that ex-ante stress-testing — systematically varying the assumptions underlying a model's fairness claims and assessing how often those claims hold — could serve as a complement to post-deployment outcome monitoring. The present framework provides a template for conducting such tests in the absence of historical deployment data.

Variable selection documentation. The proxy-bias experiment demonstrates that the choice of variables included in an AI lending model materially affects demographic outcomes, independent of whether those variables are intended to encode demographic characteristics. Governance frameworks that require lenders to document and justify their variable selection choices — specifically, to identify which variables may function as demographic proxies and to assess the magnitude of proxy loading — would directly address the mechanism identified here.

Distributional disclosure over point estimates. The standard practice of reporting a single fairness metric — such as an overall approval rate gap — provides an incomplete picture when that metric is highly sensitive to modelling assumptions. The simulation results suggest that regulatory disclosure requirements might be more informative if they required institutions to report

the sensitivity of their fairness metrics to plausible changes in model design, in addition to the point estimates themselves.

Periodic model re-evaluation under distributional shift. The static nature of the present simulation is a limitation, but it also highlights a governance gap: AI models deployed in dynamic credit markets are subject to distributional shift as borrower populations, economic conditions, and institutional practices evolve. The finding that weak fairness parameter assumptions produce adverse outcomes for SC/ST and rural borrowers suggests that model re-evaluation against updated fairness benchmarks should be a standard component of AI model governance, particularly for lending to structurally vulnerable groups.

None of these observations translates directly into a specific regulatory mandate. The appropriate institutional design for AI fairness oversight in Indian finance depends on factors outside the scope of this simulation — including the capacity of the RBI and other regulators to implement and enforce disclosure requirements, the incentive structures facing lenders, and the availability of demographic data at the borrower level. These are empirical and institutional questions that lie beyond what a simulation-based study can adjudicate.

12. Limitations

Several limitations of this study merit explicit acknowledgement.

Absence of administrative data: All results derive from synthetic data calibrated to secondary evidence. The simulation cannot validate its own assumptions against observed lending outcomes. Results should be understood as illustrative scenario estimates rather than causal evidence.

Parameter calibration uncertainty: The distributional assumptions for AI fairness parameters are motivated by the available literature but are necessarily approximations. Alternative distributional choices — for example, skewed rather than uniform distributions for AI effects — could produce different probability estimates.

Homogeneous institution assumption: The simulation does not model heterogeneity across AI-adopting institutions beyond the parameter draws. In practice, AI adoption quality varies substantially across lender types and regions in India; this heterogeneity may matter for understanding aggregate fairness outcomes.

Static model: The simulation does not model how AI adoption spreads over time, nor how fairness outcomes might evolve as models are retrained on accumulating lending data. Dynamic feedback effects — such as selection of borrowers based on model outputs — are not captured.

Binary treatment: AI adoption is treated as binary. In practice, there is substantial heterogeneity in how AI is integrated into lending decisions, ranging from automated credit scoring to AI-assisted human review.

13. Conclusion

This paper has developed a Monte Carlo simulation framework to examine the conditions under which AI-based credit scoring may reduce or amplify demographic disparities in Indian microfinance. The key finding is that AI's fairness effects are highly sensitive to both the design of the AI system and the nature of the disadvantage being addressed. Under baseline assumptions calibrated to the available secondary evidence, the simulations indicate that AI is more likely than not to reduce gender-based approval gaps (in 74.3% of simulation runs), while its effects on caste- and geography-based disparities are statistically indistinguishable from chance (approximately 51–54%). Under weak fairness assumptions, the probability of AI worsening caste and rural outcomes exceeds 50%.

The proxy-bias experiment demonstrates that variable selection is a key mechanism: lending models that incorporate stronger proxy variables correlated with protected group membership are associated with lower probabilities of improving outcomes for structurally excluded groups. The sensitivity analysis confirms that the design of AI systems — specifically, the degree to which fairness objectives are deliberately incorporated — is the primary determinant of whether AI narrows or widens caste and rural disparities.

These findings, while simulation-based, motivate several directions for future empirical research. First, the emergence of administrative datasets on AI-enabled lending decisions in India — whether through regulatory reporting requirements under the FREE-AI framework or through research partnerships with lending institutions — would allow the simulation's structural assumptions to be tested against observed outcomes. Second, natural experiment designs exploiting variation in AI adoption timing across lender types and regions could provide quasi-experimental estimates of AI's average effects on demographic lending disparities. Third, audit studies examining the variable sets actually used by AI-deploying lenders in India could directly test the proxy-bias hypotheses explored in this paper.

The paper's methodological contribution — a reproducible Monte Carlo framework for evaluating AI fairness claims under parameter uncertainty — is designed to be adapted and extended as richer data become available. The simulation code is provided in full in the Technical Appendix.

References

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56.

Blundell, R., & Stoker, T. M. (2005). Heterogeneity and aggregation. Journal of Economic Literature, 43(2), 347–391.

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

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

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

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1), 5–47.

Ghosh, S. (2013). Macroeconomic uncertainty and credit access: Evidence from Indian states. Economic & Political Weekly, 48(3).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. AEA Papers and Proceedings, 108, 22–27.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics, 29(3), 305–325.

Marda, V., & Narayan, S. (2020). Data in New Delhi's predictive policing system. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 317–324).

Reserve Bank of India. (2017, October 4). Non-Banking Financial Company – Peer to Peer Lending Platform Directions, 2017. RBI.

Reserve Bank of India. (2022a, September 2). Guidelines on digital lending. RBI.

Reserve Bank of India. (2025). Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI) in the financial sector. RBI.

Sane, R., & Halan, M. (2017). Misled and mis-sold: Financial misbehaviour in retail banks? Journal of Comparative Economics, 45(3), 429–444.

Women's World Banking. (2022, March 10). In a world of gender bias, Lendingkart's AI-based credit model stands apart. Women's World Banking. https://www.womensworldbanking.org/insights/in-a-world-of-gender-bias-lendingkarts-ai-based-credit-model-stands-apart/

Technical Appendix

A1. Full Parameter Specifications

All simulation parameters are documented in Table 1 of the main text. The Python code appendix (A2) provides the complete implementation. Random seeds are fixed for reproducibility: np.random.seed(42) is set globally; each iteration uses a deterministic seed of BASE_SEED + iteration_index.

A2. Python Code

The following code reproduces all simulations, figures, and tables reported in this paper. Dependencies: numpy, pandas, scipy, statsmodels, matplotlib (all standard in Anaconda/pip distributions).

```
"""
Monte Carlo Simulation: AI and Demographic Disparities in Indian Microfinance
Replication code for: When Does AI Reduce Bias?
Requirements: numpy, pandas, scipy, statsmodels, matplotlib
"""


import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import warnings, os
warnings.filterwarnings("ignore")
np.random.seed(42)

N_BORROWERS = 3000
N_ITER      = 1000
AI_SHARE    = 0.208   # RBI FREE-AI: 20.8% adoption rate
```

```
SEED_BASE   = 42
os.makedirs("outputs", exist_ok=True)

def generate_borrowers(n=N_BORROWERS, seed=None):
    rng = np.random.default_rng(seed)
    female   = rng.binomial(1, 0.48, n)
    scst     = rng.binomial(1, 0.22, n)
    rural    = rng.binomial(1, 0.62, n)
    income   = rng.integers(1, 11, n)
    credit   = rng.binomial(1, 0.62, n)
    employed = rng.binomial(1, 0.71, n)
    ai_inst  = rng.binomial(1, AI_SHARE, n)
    return pd.DataFrame({
        "female":female, "scst":scst, "rural":rural,
        "income_dec":income, "credit_hist":credit,
        "employed":employed, "ai":ai_inst,
        "income_scaled":(income - 5.5) / 4.5
    })

def run_iter(seed, ag=None, asc=None, ar=None, proxy_mode="none"):
    rng = np.random.default_rng(seed)
    # Draw AI fairness parameters
    if ag  is None: ag  = rng.uniform(0.05, 0.20)
    if asc is None: asc = rng.uniform(-0.05, 0.10)
    if ar  is None: ar  = rng.uniform(-0.05, 0.08)
    df = generate_borrowers(seed=seed)
    n  = len(df)
    # Baseline penalties (calibrated to literature)
    b_scst = -0.20; b_rural = -0.16
    if proxy_mode == "moderate":
        b_scst  -= rng.uniform(0.02, 0.06)
        b_rural -= rng.uniform(0.02, 0.05)
    elif proxy_mode == "strong":
        b_scst  -= rng.uniform(0.06, 0.14)
        b_rural -= rng.uniform(0.05, 0.12)
    lp = (-0.30 - 0.14*df["female"] + b_scst*df["scst"]
          + b_rural*df["rural"] + 0.12*df["income_scaled"]
          + 0.38*df["credit_hist"] + 0.18*df["employed"]
          + ag*df["ai"]*df["female"]
          + asc*df["ai"]*df["scst"]
          + ar*df["ai"]*df["rural"]
          + rng.normal(0, 0.22, n))
    p  = 1 / (1 + np.exp(-lp))
    df["approved"] = rng.binomial(1, p, n)
    # Compute approval gap change
    def gap_delta(col):
        a = df[df["ai"]==1]; b_ = df[df["ai"]==0]
        def d(x):
            return x[x[col]==1]["approved"].mean() -
x[x[col]==0]["approved"].mean()
        return d(a) - d(b_)
    # OLS regression
    X = sm.add_constant(df[["ai","female","scst","rural",
                            "income_scaled","credit_hist","employed"]])
    X["ai_female"] = df["ai"] * df["female"]
    X["ai_scst"]   = df["ai"] * df["scst"]
    X["ai_rural"]  = df["ai"] * df["rural"]
    m = sm.OLS(df["approved"], X).fit()
    return {
        "gd":  gap_delta("female"),
        "sd":  gap_delta("scst"),
        "rd":  gap_delta("rural"),
```

```
        "c_af": m.params.get("ai_female", np.nan),
        "c_as": m.params.get("ai_scst",   np.nan),
        "c_ar": m.params.get("ai_rural",  np.nan),
    }

# Baseline MC
res = [run_iter(SEED_BASE + i) for i in range(N_ITER)]
mc  = pd.DataFrame(res)
mc.to_csv("outputs/mc_baseline.csv", index=False)

# Scenarios
sc1 = [run_iter(10000+i, ag=np.random.uniform(0.15,0.25),
                asc=np.random.uniform(0.05,0.15),
                ar=np.random.uniform(0.04,0.12)) for i in range(N_ITER)]
sc2 = [run_iter(20000+i, ag=np.random.uniform(-0.02,0.06),
                asc=np.random.uniform(-0.10,0.01),
                ar=np.random.uniform(-0.10,0.00)) for i in range(N_ITER)]
sc1_df = pd.DataFrame(sc1); sc2_df = pd.DataFrame(sc2)

# Proxy experiment
for mode in ["none", "moderate", "strong"]:
    px = [run_iter(50000+i, proxy_mode=mode) for i in range(N_ITER)]
    pd.DataFrame(px).to_csv(f"outputs/proxy_{mode}.csv", index=False)

# Print summary
print("Baseline P(AI reduces gap):")
print(f"  Gender: {(mc['gd']>0).mean():.3f}")
print(f"  SC/ST:  {(mc['sd']>0).mean():.3f}")
print(f"  Rural:  {(mc['rd']>0).mean():.3f}")
```

This code, along with the figure-generation routines, is available in full as a supplementary file submitted alongside this paper. All random seeds are fixed for exact reproducibility.