

Generous to Men or Harsh to Women?

Experimentally Unpacking Gender Bias in Lending

Youngjoo Jung*

Job Market Paper | December 2025

[Click here for the latest version](#)

Abstract

In low-income economies with limited financing options, loan officers act as gatekeepers to capital, and their biases can distort capital allocation. I study gender bias in small business lending in Egypt by presenting 720 loan officers with past loan applications where I randomized applicant names. The same application is approved 6.2% more often if it has a male name, with the gap concentrated in low-quality loans that later defaulted, had late repayment history, or low credit scores. This shows that the bias manifests as generosity toward men rather than harshness toward women. Separating out bias in this way is critical to understanding its implications and how best to combat it. I test different strategies to decrease bias. Implicit Association Test feedback does not appear effective, while performance pay leads loan officers to increase screening effort and eliminates the bias. Combining both treatments also closes the gap without increasing screening effort. However, when officers receive Generative AI assistance, they are more likely to follow approval suggestions for low-quality loans with male names.

Keywords: Gender bias, Loan officer, Entrepreneurship, Performance pay, AI, Lending

JELClassification: D9, G4, G5, O1

*Youngjoo Jung, yj27@illinois.edu, University of Illinois at Urbana-Champaign. Approval from University of Illinois Human Subjects Committee (IRB 24-0796) and American University in Cairo Human Subjects Committee (IRB 2023-2024-235). This project was registered on the AEA RCT Registry (AEARCTR-0015025). I greatly appreciate my advisor, Adam Osman, for his sincere support. I am grateful for Richard Akresh, Julia Fonseca, Lena Song, Marique Kleemans, Dan Bernhardt, Joshua Shea, Hoyoung Yoo, John List, Robert Metcalfe, Marianne Bertrand, Michelle Brock, Ralph De Hass, Kareem Haggag, Muhammad Meki, Alex Chan, Ben Roth, Natalia Rigol, Reshma Hussam, Bennedek Kurdi, Maria Loumiodi, Leora Klapper, Jules Gazeaud, Carolyn Pelnik, Carolyn Tsao, Diego Ubfal, Sophie Shive, Silvia Prina, Sanghee Mun, Chan Kim, Yeonggyu Yun, Abigail Stocker, William Jeziorski, Sullam Jeoung, and participants at UIUC AMRL, job market seminar, AFE 2025, FMA Doctoral Consortium, EGSC 2025, KAEA Job Market Conference, University of Chicago experimental economics working group, and Accelerating Growth for Women Entrepreneurs Research Conference 2025 for their superb advice. I would also like to thank the Weiss Fund for their funding. For project management and field support, I am grateful to Family Development Foundation and J-PAL MENA staff, especially Mahmoud Hussein and Ahmed Gawish. All errors are those of the author.

1 Introduction

In many low-income economies, where lenders are scarce and small businesses face limited financing options, loan officers act as gatekeepers to capital (Rigol and Roth, 2021). Their bias not only limits credit access for individual businesses but can also contribute to capital misallocation and productivity loss (Hsieh and Klenow 2009; Moll 2014; Hsieh et al. 2019). The persistent gender gap in entrepreneurial lending (Demirguc-Kunt et al., 2018; Elam et al., 2021; Ubfal, 2024) suggests potential gender bias in small business loan approval, but prior work has shown limited evidence of bias (Brock and De Haas, 2023; Ayalew et al., 2023). Moreover, it remains unclear whether bias stems from generosity toward men—approving unqualified men in low-quality loans—harshness toward women—rejecting qualified women in high-quality loans—or both.

Disentangling generosity toward men and harshness toward women is empirically challenging. Identifying gender bias requires isolating gender as the only varying attribute. Unpacking the nature of bias goes a step further: it requires observing how gendered decisions vary across applicants with differing quality. This is rarely feasible in observational settings, and even experimental studies to date have typically not unpacked this. Yet, the distinction is crucial. While both biases reduce lenders’ profits, each entails different policy implications. Generosity undermines financial stability, particularly during negative economic shocks. Conversely, harshness restricts qualified borrowers’ access to finance and hurts the economy at large by keeping qualified women’s businesses smaller than optimal. Given the type of the bias, debiasing policies need to be tailored to reduce lending costs, especially in low-income economies with high information asymmetries.

To overcome these challenges, I implement an incentivized¹ framed field experiment (Harrison and List, 2004; Cole et al., 2015) with 720 microfinance loan officers in Egypt. I provide examples of previous small business loan applications and ask officers if they would approve or deny the applicant with randomized names. To test potential solutions, officers are randomly assigned to a control or one of three debiasing treatment groups: (1) a nudge using Implicit Association Test (IAT) feedback² targeting intrinsic motivation, (2) higher performance pay targeting extrinsic motivation, (3) or a combined intervention with both.

I find strong evidence of gender bias. The same application is 6.2% more likely to be approved when randomly assigned a male name. The nature of this gender gap stems mainly from generosity toward men, as the pro-male gap is concentrated among low-quality loans. This pattern holds across three measures of quality: First, for ex-post defaulted loans, the gender gap is 14.7%, compared to only 1.2% for ex-post repaid loans. Second, for loans with a history of late repayment, the gap is 21.6%, versus 3.4% for those without. Third, the gap is 16.8% for low credit score loans, but only 0.8% for high credit score loans.

¹All officers receive performance pay based on the ex-post loan outcomes: bonus for approving repaid loans or rejecting defaulted loans, and penalties for approving defaulted loans or rejecting repaid loans.

²The Gender–Career IAT measures unconscious bias by recording how quickly individuals associate men with careers and women with family. The feedback presents the score and its interpretation.

Turning to the debiasing interventions, a nudge using IAT feedback does not appear effective, whereas both higher performance pay and the combined intervention close the gender gap but through different mechanisms. Both treatments improve screening performance for male-name loans, with higher performance pay increasing screening time, whereas the combined intervention improves screening performance without extending screening time. A back-of-the-envelope calculation suggests that generosity toward men costs the bank approximately EGP 3.6 billion (USD 72.9 million) per year, and that increasing performance pay could raise annual net profits by roughly EGP 3.2 billion (USD 64.3 million).

There is no heterogeneity in the gender gap in loan approval by officers' gender or IAT score. However, less-experienced or lower-educated officers display larger bias. Similarly, I find no heterogeneity in the debiasing effects of higher performance pay by officers' gender or IAT score, but the effects are stronger among less-experienced or lower-educated officers.

To understand the mechanisms behind generosity toward men and the debiasing effect of higher performance pay, I develop a Bayesian theoretical framework of loan officers' lending decisions that incorporates gender bias. Officers hold prior beliefs and choose optimal screening effort to increase signal precision under performance pay tied to repayment outcomes. The model shows that generosity toward men arises when officers either derive taste-based utility from lending to men or hold optimistic priors about their repayment probability, but not disutility from lending to women or pessimistic priors about women's repayment. When officers are generous toward men, higher performance pay increases their screening effort and raises the approval threshold for male applicants by reducing the influence of taste-based preferences or optimistic priors on lending decisions. Consistent with these predictions, the experiment shows that higher performance pay increases screening time and closes the gender gap by improving screening performance for male-name loans.

The experiment's second round moves beyond traditional methods to test AI assistance as a potential solution. I analyze whether a generative AI like ChatGPT can improve loan officers' screening performance, especially in low-income economies contexts where resources are limited, officers have fewer skills, and defaulted loan rates are high (Eyraud et al. 2021). In this round, officers evaluate a new set of loans without access to their first-round results. They receive recommendations from a custom AI chatbot I built with GPT-4o on whether to approve or reject each applicant, accompanied by two justifications. Stratified by their initial treatment, officers are further randomized into one of three second-round groups: (1) a control group with no AI assistance, (2) a non-generative AI group that passively views chatbot suggestions, similar to access to a better algorithm, and (3) a generative AI group that actively interacts with the chatbot through follow-up questions, similar to ChatGPT.

I find that officers using either generative or non-generative AI show better screening performance relative to those without AI. However, the communication feature of generative AI provides no additional benefit.

Instead, it may reveal a channel through which existing human biases manifest when the generative AI makes erroneous recommendations aligning with officers pre-existing bias. When the AI recommends approval for ex-post defaulted loans, generative AI users are more likely to follow these recommendations for male-named applications, suggesting that generosity toward men persists even under AI guidance. Supporting this, female-named portfolios receive more questions via the chatbot, including more expressions of disagreement with the AI’s approval recommendation for ex-post defaulted loans. By contrast, when generative AI advises approving repaid loans or rejecting defaulted loans, no differences appear by applicant gender.

I contribute to the literature on biased decision-making by being, to the best of my knowledge, the first to experimentally separate generosity and harshness. While theoretical decomposition of bias into generosity and harshness was proposed decades ago in psychology (Glick and Fiske, 1996), the broader economics literature—across labor (Altonji and Blank, 1999; Bertrand and Mullainathan, 2004; De Mel et al., 2009; Goldin, 2014; Blau and Kahn, 2017; Glover et al., 2017; Baert, 2018; Neumark, 2018; Agan and Starr, 2018; Mocanu, 2022; Egan et al., 2022; Osman et al., 2023), housing (Hanson et al., 2016; Flage, 2018; Bao, 2024), teaching (Mengel et al., 2019; Arnold et al., 2020), judicial decisions (Arnold et al., 2018), and lending (Alibhai et al., 2019; Brock and De Haas, 2023; Ayalew et al., 2023; Montoya et al., 2024; Bartös et al., 2024)³—has generally examined whether discrimination exists, rather than whether it arises from generosity toward in-group or harshness toward out-group. Some studies find asymmetric discrimination (Hanson et al., 2016; Sarsons, 2017; Egan et al., 2022; Hebert et al., 2024), yet without ground truth on applicant quality, these designs are not likely to distinguish generosity from harshness under stochastic outcomes. In the lending context, by accessing loan quality, I distinguish generosity from harshness and show that the observed gender bias is primarily driven by generosity toward men.

I also contribute to the literature on loan officers’ gender bias by providing the first experimental evidence of gender bias in small business loan approvals. While previous observational studies report gender bias in lending (Carter et al., 2007; Muravyev et al., 2009; Stefani and Vacca, 2013; Beck et al., 2018), experimental evidence remains scarce and mixed. Prior experimental studies document gender bias in officers’ evaluations of business proposals in Uganda (Bartös et al., 2024), loan allocation task in Turkey (Alibhai et al., 2019), consumer loan approvals in Chile (Montoya et al., 2024), and early-stage equity financing in France (Hebert, 2023), but none in business loan approvals. This is partly because they examine settings where gender

³This includes a large body of observational research examining bias in credit markets along different dimensions, including gender (Carter et al., 2007; Muravyev et al., 2009; Barasinska and Schäfer, 2010; Bellucci et al., 2010; Agier and Szafarz, 2013; Alesina et al., 2013; Stefani and Vacca, 2013; Mascia and Rossi, 2017; Beck et al., 2018; Andreeva and Matuszyk, 2019; Hebert, 2023; Hebert et al., 2024), race (Berkovec et al., 1998; Cavalluzzo and Cavalluzzo, 1998; Cavalluzzo et al., 2002; Blanchflower et al., 2003; Charles and Hurst, 2002; Han, 2004; Charles et al., 2008; Ross et al., 2008; Cohen-Cole, 2011; Pope and Sydnor, 2011; Deku et al., 2016; Sarsons, 2017; Bayer et al., 2018; Arnold et al., 2018; Fuster et al., 2022; Arnold et al., 2022; Butler et al., 2023; Frame et al., 2025), cultural proximity (Fisman et al., 2017), social networks (Haselmann et al., 2018), and age or immigrant status (Dobbie et al., 2021).

bias is less likely to emerge. For example, (Brock and De Haas, 2023) found gender bias in collateral requirements in Turkey’s banking sector but no bias in loan approval decisions—likely because their sample focused on formal bank lending, which tends to involve high-quality applications, making it harder to detect bias stemming from generosity toward men in low-quality applications. Similarly, Ayalew et al. (2023), using data from a national business competition in Ethiopia, found no gender bias, possibly because the applicants were selective, high-performing entrepreneurs and the evaluators were highly experienced officers. In contrast, I focus on loans with mid-range credit scores, where loan officers’ discretion is likely greater and gender bias more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024), providing evidence of gender bias in small-business loan approval.

This paper also contributes to the literature on strategies for mitigating gender bias by comparing intrinsic, extrinsic, and combined approaches within the same setting. Most existing studies focus on either extrinsic or intrinsic methods alone. Extrinsic incentives such as performance pay have a complex relationship with bias, depending on the objectivity of the evaluation process (Heywood and O’Halloran, 2005; Heywood and Parent, 2012; Marinescu et al., 2018; Mejia and Parker, 2021). Intrinsic approaches—including greater exposure to the out-group (Beaman et al., 2009), information interventions (Bohren et al., 2019; Montoya et al., 2024), education (Amaral et al., 2021), or implicit bias awareness (Alesina et al., 2024)—have shown some effectiveness in reducing bias. Building on this literature, I show that higher performance pay can serve as an effective debiasing tool for loan officers where performance is objectively measured by ex-post repayment outcomes. By contrast, a behavioral nudge via IAT feedback alone appears to have insufficient impact. Moreover, I provide the first experimental evidence that combining higher performance pay with IAT feedback produces a more complex interaction in debiasing rather than a simple additive effect.

I further contribute to the literature on AI use in credit markets and biased decision-making by showing that even when generative AI provides gender-neutral recommendations, biased user behavior can produce gender-biased lending outcomes. Existing work has primarily focused on bias embedded in non-generative AI models, such as algorithmic and machine-learning-based credit scoring systems (Buolamwini and Gebru, 2018; Berente et al., 2021; Fu et al., 2021; Angwin et al., 2022; Bartlett et al., 2022; Fuster et al., 2022; Dastin, 2022; Agarwal et al., 2023; Hurlin et al., 2024), or on bias within generative AI systems themselves (Bowen III et al., 2024; Etgar et al., 2024; Cook and Kazinnik, 2025). In many real-world lending environments, however, humans remain the final decision-makers, with AI serving as an assistive tool rather than a substitute for human judgment. While Hu et al. (2025) show that human bias can distort the use of non-generative AI, there is no prior empirical evidence on whether and how such bias manifests under generative AI assistance in financial decision-making. Holding generative AI recommendations fixed and varying only applicant names, I show that loan officers selectively follow AI approval recommendations more

often for male-named applications. I further document the mechanism underlying this pattern by analyzing the communication between generative AI and loan officers.

Lastly, I contribute to the literature on human adherence to AI recommendations and AI-assisted worker performance by comparing generative and non-generative AI within the same lending environment. Prior work shows that human overrides of algorithmic recommendations can attenuate the gains from AI assistance, including in judges’ sentencing decisions (Stevenson and Doleac, 2024), bail decisions (Angelova et al., 2023), radiology (Agan et al., 2023; Yu et al., 2024), hiring (Fumagalli et al., 2022; Dargnies et al., 2024), and theoretical models of human–AI collaboration (Grand-Clément and Pauphilet, 2024; Agarwal et al., 2025). Foroughifar et al. 2025 shows that uncertainty disclosure in algorithmic recommendations affects human adherence. Recent work documents mixed effects of generative AI assistance on work outcomes, including positive effects for customer support agents (Brynjolfsson et al., 2025), diagnostic reasoning by physicians (Goh et al., 2024), and writing tasks (Noy and Zhang, 2023), but mixed findings in consulting work (Dell’Acqua et al., 2023) and no average business performance gains among entrepreneurs, with substantial heterogeneity across users (Otis et al., 2023). None of these studies, however, compare performance under non-generative and generative AI assistance. By comparing generative and non-generative AI within the same lending environment, I show that loan officers selectively override generative AI recommendations in a gender-biased manner, whereas no such pattern emerges when officers receive non-generative AI recommendations. A plausible interpretation is that non-generative AI, which provides recommendations without interactive explanations, is more likely to be followed mechanically, whereas generative AI encourages longer screening time and active engagement, increasing opportunities for discretionary judgment and, consequently, biased overrides.

The paper is structured as follows. Section 2 provides the context and theoretical framework. Section 3 introduces the experimental design and data. Section 4 presents the main results. Section 5 discusses heterogeneous effects, and concluding remarks are offered in Section 6.

2 Context and Theoretical Framework

The experiment is conducted in Egypt, which offers a suitable setting for testing gender bias and the nature of bias. Discrimination against women is rampant (Osman et al., 2023), and the impacts of microcredit are shown to be greater for women compared to men (Crépon et al., 2024). Evidence from the World Values Survey (Haerpfer et al., 2022) further highlights the potential of gender bias in Egypt. Figure A1 presents comparative responses to six gender-related questions from the 2017 survey in Egypt and the United States. For example, 36.8% of Egyptian respondents strongly agreed with the statement, “On the whole, men make

better business executives than women do,” compared to just 2.2% in the U.S. The contrast is even starker for the belief that “When jobs are scarce, men should have more right to a job than women,” with 67.4% of Egyptians strongly agreeing versus 0% of Americans. These patterns suggest a strong presence of gender bias in Egypt, potentially influencing lending decisions.

The experiment targets loan officers at microfinance institutions (MFIs) in Egypt. Between November 2024 and May 2025, 720 officers from 29 MFIs participated in the study. Figure 1 maps the five governorates where the experiment took place: Aswan, Asyut, Luxor, Minya, and Qena. Table A1 provides a list of the MFIs and the number of participating officers from each. The largest participating MFI is the Family Development Foundation, based in Aswan and Luxor, and the experiment uses administrative loan application data from this institution. The participating officers are gender-balanced, relatively young (average age 34.6), and well-educated (16.2 years of schooling), with an average of 6.3 years of experience as loan officers. Loan officers used to approve 67.3% of the loans assigned to them in the reality. They exhibit moderate implicit gender bias, with a mean Gender-Career IAT score of 0.37—suggesting a medium or stronger association of men with career and women with family. Regarding risk preferences, officers report an average score of 6.44 on a 0–10 scale of risk preference and indicate they would invest 52.3% of an endowment in a fair 50/50 investment choice, suggesting they are not generally risk-averse. Full details are provided in Column (1) of Table 1.

In Egypt, loan officers typically receive a small base salary supplemented by performance bonuses tied to loan repayment outcomes, but face non-monetary penalties, such as disadvantages in promotion, for poor loan repayment performance. Reflecting this structure, all officers receive performance pay based on ex-post loan outcomes in the experiment (Cole et al., 2015; Brock and De Haas, 2023): bonuses for approving repaid loans or rejecting defaulted loans, and penalties for approving defaulted loans or rejecting repaid loans, with a heavier penalty for approving defaulted loans. This design captures the asymmetry in real-world incentive structures, where the costs of default weigh more heavily than the benefits of repayment.

The experiment mainly targets loan officers’ taste-based bias (Becker, 1957), while also acknowledging the existence of statistical discrimination (Phelps, 1972; Aigner and Cain, 1977; Bohren et al., 2023). Even with identical applications, if male businesses operate in more favorable environments, officers may be more likely to approve men. Field evidence further suggests that both industries and officers’ priors may be gender-segregated, as officers rarely visit businesses run by the other gender alone. To mitigate potential statistical discrimination, officers are informed that repayment outcomes are not differ across genders (60%) in the loan sample on average, and the loan set is drawn from industries with balanced gender composition⁴ using administrative data from the Family Development Foundation. However, since beliefs adjust gradually

⁴Industries in which at least 30% of borrowers are men or women.

(Bohren et al., 2019), a one-time information intervention may be insufficient to fully override prior beliefs. Thus, some degree of statistical discrimination may still operate.⁵

The goal of this experiment is to identify both the existence and nature of gender bias in loan approval decisions and to test the effectiveness of debiasing interventions. Before describing the experiment, I present a theoretical framework to formalize how officers' taste-based preferences or biased prior beliefs about repayment probability can result in either generosity toward men or harshness toward women in lending decisions. I also illustrate how increasing performance pay can change screening effort and alter biased lending decisions.

2.1 Model Primitives

2.1.1 Loan Officer

Risk-neutral loan officers receive performance pay based on the accuracy of their approval decision by the loan's ex-post outcome. To capture the fact that the stakes from defaults are typically greater than those from repayments, I assign two stake parameters with $\ell_1 > \ell_2 > 0$ where $\pi \geq 1$ represents the strength of performance pay.

- Approve: $+\pi\ell_2$ if repaid, $-\pi\ell_1$ if defaulted.
- Reject: $+\pi\ell_1$ if defaulted, $-\pi\ell_2$ if repaid.

The officer may have a constant taste-based (dis)utility, d , which depends on the applicant's gender, G_j and the approval decision D_j , and is captured by non-negative parameters: δ_M denotes the utility from lending to men, and δ_F the disutility from lending to women.

$$d = \begin{cases} \delta_M & \text{if } G_j = M \text{ and } D_j = 1 \\ -\delta_F & \text{if } G_j = F \text{ and } D_j = 1 \\ 0 & \text{otherwise} \end{cases}$$

2.1.2 Loan Applicant (j)

Each loan applicant's project has a true, unobservable revenue, r . The loan amount is normalized to 1, so a loan is repaid if $r \geq 1$ and defaulted if $r < 1$. The officer's prior belief about revenue r for an applicant follows a normal distribution with a gender-specific mean m_{G_j} and a common variance σ_r^2 .

$$r \sim N(m_{G_j}, \sigma_r^2)$$

⁵Loan officers may not fully trust that repayment outcomes do not differ by gender in the loan sample and may perceive the selected industries as male-dominated. Unfortunately, I do not observe whether participants were aware of it during the experiment.

Officers may hold more favorable prior beliefs toward men, and loan applications are on average profitable (i.e., $m > 1$). Therefore, I assume $m_M \geq m_F > 1$. Moreover, to reflect the high uncertainty in microfinance lending (e.g., informal and micro-business contexts), I assume the prior variance σ_r^2 is sufficiently large.

2.2 Information Structure and Decision Rule

The officer cannot observe r . Instead, they choose an effort e_j and receive a noisy signal, $s = r + \epsilon$. The error term ϵ follows a normal distribution with mean zero and variance $\sigma_\epsilon^2(e_j)$, which decreases with effort.

$$\epsilon \sim N(0, \sigma_\epsilon^2(e_j)), \quad \text{where} \quad \sigma_\epsilon^2(e_j) = \frac{1}{e_j}$$

With the signal s , the officer uses Bayesian updating to form a posterior belief about the loan's revenue:

$$r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2), \quad \text{where} \quad \hat{m}_{G_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

The officer's belief about the loan being repaid, $p(s, e_j, G_j)$, is the posterior probability that loan revenue $r \geq 1$, expressed using the cumulative distribution function of the standard normal distribution Φ .

$$p(s, e_j, G_j) = P(r \geq 1 \mid s, e_j, G_j) = \Phi \left(\frac{(m_{G_j} - 1) + e_j \sigma_r^2 (s - 1)}{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)}} \right)$$

The officer approves a loan if the expected utility of approval exceeds that of rejection. Since $p(s, e_j, G_j)$ is monotonically increasing in the signal s , this is equivalent to approving the loan when $s \geq s_{G_j}^*$.

$$s_{G_j}^*(e_j, \pi, d) = 1 + \frac{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)} \cdot \Phi^{-1}(k_{G_j}^*) - (m_{G_j} - 1)}{e_j \sigma_r^2}$$

where $k^*(G_j = M) = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1 + \ell_2)} \leq k^*(G_j = F) = \frac{2\pi\ell_1 + \delta_F}{2\pi(\ell_1 + \ell_2)}$. The full derivations are provided in Appendix E.

2.3 The Officer's Optimal Effort Choice

The officer chooses an optimal effort level e_j to maximize expected utility subject to the convex cost ce^2 .

$$\max_{e_j \geq 0} U(e_j; G_j, \pi, d) = \underbrace{\int_{-\infty}^{s_{G_j}^*(e, \pi, d)} V_R(s, e, G, \pi) g(s \mid e, G) ds}_{\text{reject region}} + \underbrace{\int_{s_{G_j}^*(e, \pi, d)}^{\infty} V_A(s, e, G, \pi, d) g(s \mid e, G) ds}_{\text{approve region}} - ce^2.$$

where $V_R(s, e, G, \pi) = \pi[\ell_1 - (\ell_1 + \ell_2)p(s, e, G)]$, $V_A(s, e, G, \pi, d) = \pi[-\ell_1 + (\ell_1 + \ell_2)p(s, e, G)] + d(G)$,

$g(s | e, G)$ denotes the probability density function of the signal.

First-Order Condition can be written as three terms: (i) higher effort improves decision accuracy, increasing monetary benefits (A term); (ii) higher effort shifts the approval threshold, changing the relative weight of taste-based bias (B term); and (iii) the marginal cost of effort (see Appendix E for full derivations).

$$\frac{dU}{de} = \underbrace{\pi A(e; k_G^*(\pi, d(G)))}_{\text{(i) direct monetary return}} + \underbrace{d(G) B(e; k_G^*(\pi, d(G)))}_{\text{(ii) taste-based bias return}} - \underbrace{2ce}_{\text{(iii) marginal cost}}$$

2.4 Testable Predictions

The model formalizes how taste-based bias or priors about repayment likelihood can shape generosity toward men and harshness toward women in approval thresholds. It also shows how higher performance pay affects both screening effort and approval thresholds, yielding testable predictions. Proofs are in Appendix E.

Proposition 1. *If effort e is sufficiently large⁶, the presence of taste-based bias ($\delta_M > 0$ or $-\delta_F < 0$) or biased prior belief ($m_M > m_F$) leads to a lower approval threshold for male applicants ($s_M^* < s_F^*$).*

Corollary 1. *Based on Proposition 1, the model predicts two distinct patterns by the type of bias:*

- (a) *If $\delta_M > 0$ and $\delta_F = 0$, or if $m_M > m_{unbiased} = m_F$, then $s_M^* < s_{unbiased}^* \approx s_F^*$. As a result, unqualified male applicants with $s \in (s_M^*, s_{unbiased}^*)$ are approved, indicating generosity toward men.*
- (b) *If $\delta_M = 0$ and $-\delta_F < 0$, or if $m_M = m_{unbiased} > m_F$, then $s_M^* \approx s_{unbiased}^* < s_F^*$. As a result, qualified female applicants with $s \in (s_{unbiased}^*, s_F^*)$ are rejected, reflecting harshness toward women.*

Proposition 2. *For a sufficiently large increase in performance pay (π), the officer's optimal screening effort rises.*

Testable Prediction 1: Higher performance pay and screening effort

Loan officers with the higher performance pay treatment are expected to spend more time on loan screening relative to officers in the control group.

Proposition 3. *If either δ_M or m_M is sufficiently large, higher performance pay raises the approval threshold for male applicants (s_M^*).*

⁶In microfinance settings, loan officers face noisy information and high uncertainty. This corresponds to a high prior variance σ_r regarding repayment likelihood, which is likely to result in large optimal effort levels to get more precise signal.

Testable Prediction 2: Higher performance pay and male applicants’ approval

If the pro-male gender gap in loan approval is driven by generosity toward men, higher performance pay is expected to increase the approval threshold for male applicants.

3 Experimental Design and Data

3.1 Experimental Design

The experiment is introduced as an assessment of officers’ ability to screen for defaulted loans to prevent them from realizing its true purpose—measuring gender bias. The field team coordinates with MFIs and schedules visits for data collection. Experiments are conducted individually, with one enumerator assigned per officer. When MFIs lack sufficient space, nearby rented facilities are used. All activities are administered using SurveyCTO. Loan application information is provided in printed form to replicate typical screening practices. Officers assigned to the AI treatments in the second-round receive laptops to facilitate AI-assistance.

The experiment begins with a baseline survey collecting data on demographics, work characteristics, cognitive ability, risk preferences, psychometrics, and social desirability. Officers then complete a Gender-Career Implicit Association Test (IAT) adapted for SurveyCTO from the publicly available version⁷. Next, officers complete lending decision tasks in two rounds, with group assignments randomized separately for each round. Afterward, they answer explicit bias questions from the World Values Survey (Haerpfer et al., 2022) and questions related to AI usage. The experiment then ends. Details of the questionnaires are provided in Appendix D. On average, the total duration of the experiment was 50 minutes and no participants dropped out during the course of the experiment.

Total compensation is based on the average accuracy of loan decisions in the first and second rounds. Officers receive a base payment of 400 EGP. They earn 10 EGP for approving a loan that was ex-post repaid and 15 EGP for rejecting a loan that was ex-post defaulted. Conversely, they receive a penalty of 15 EGP for approving an ex-post defaulted loan and 10 EGP for rejecting an ex-post repaid loan. This asymmetric incentive structure reflects the real-world context, where loan officers face greater penalties for approving bad loans than for rejecting good ones. Full details of the incentive scheme are provided in Table A2. Given that the median monthly income of participating officers is approximately 6,000 EGP, the rewards for the 50 minutes experiment are substantial: depending on their decision accuracy, officers could earn between 280 EGP and 520 EGP in total compensation, and they actually received an average reward of 417 EGP.

⁷<https://implicit.harvard.edu/implicit/>. Following standard IAT procedures (Greenwald et al., 1998, 2003; Lane et al., 2007), trials shorter than 300 milliseconds or longer than 10,000 milliseconds are excluded, and the seven-block structure is applied.

3.1.1 Loan Portfolios

I use 50 previous approved loan applications⁸ in the experiment, selected based on credit scores and industries. Given the limited evidence in the literature regarding the experimental presence of gender bias in small business loan approvals (Brock and De Haas, 2023; Ayalew et al., 2023), I focus on cases where loan officers have greater discretion and where gender bias is more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024). In the Egyptian credit scoring system, there are five tiers: excellent (above 750), very good (701–750), satisfactory (626–700), not satisfactory (520–625), and bad (below 520). Also, there are many small businesses operate informally and do not have credit scores due to the limited coverage of the credit system. Then, I exclude loans with very good credit scores (above 750) and bad scores (below 520), as these represent clearly high- or low-quality applicants where gender bias is less likely to occur. I instead select applications with credit scores between 520 and 750 or with no credit score. Next, to reduce the extent to which gender provides meaningful prior information, I restrict the sample to industries where each gender represents at least 30% of borrowers, as gender composition may otherwise affect screening standards (Mocanu, 2022). The selected industries include agriculture, supermarkets, grocery stores, clothing shops, fish sale shops, poultry, and livestock.

The ratio of ex-post repaid to defaulted loans was designed to mirror real-world approval behavior and provide sufficient variation to capture generosity and harshness. Although precise approval rate data from the Family Development Foundation (which provided the loan portfolios) were unavailable, discussions with a manager indicated typical approval rates of around 70%. Consistently, in the baseline, officers approved 67.3% of assigned applications on average in their real-world workplace. Reflecting this, 60% of the 50 loans used in the experiment were ex-post repaid and 40% were defaulted, making the optimal approval rate approximately 60%. Officers are informed that the repayment rate in the experiment is set at 60% on average, though it may vary across individuals.

To isolate gender as the only varying attribute when officers screen multiple loans, it is crucial to control the composition of loan bundles, as the evaluation of a loan may differ depending on the other loans presented alongside it. To address this, I group the 50 loans into five bundles, each containing 10 loan portfolios with 60% repaid loan rate: 3 ex-post repaid male-name loans, 3 ex-post repaid female-name loans, 2 ex-post defaulted male-name loans, and 2 ex-post defaulted female-name loans. I then create Bundles 6 through 10 by duplicating Bundles 1 through 5, but swapping each applicant’s name with suggesting the opposite gender. For example, Bundles 1 and 6 contain the same 10 loan applications in the same order, but names are switched (e.g., “Noha Mahmoud” in Bundle 1 becomes “Ahmed Mahmoud” in Bundle 6). This pairing

⁸Data on rejected loans were unavailable and thus not included in the experiment. With an approval rate of about 70%, roughly 30% of applications are not captured in the analysis.

process is repeated for Bundles 2 and 7, 3 and 8, 4 and 9, and 5 and 10. While gender is not explicitly stated, fictitious first names, selected from commonly used Egyptian names, implicitly indicate the applicant’s gender. The detailed sample selection procedure is provided in Appendix B.

The loan portfolios contain standard financial information used in real-world loan screenings in the partner MFI, including name, age, number of guarantors, monthly income, business registration status, business start year, location, number of employees, industry, loan size, loan tenure, interest rate, number of previous loans, number of delayed days on the most recent loan, loan size and interest rate of the most recent loan, and credit score. Similar to real-world screening, officers review these printed portfolios by bundle and make their decisions. Each page contains 5 portfolios on one side, so one sheet of paper includes a total of 10 portfolios on both sides. Figure A2 displays a pseudo example of a loan application used in the experiment.

3.1.2 Randomization

I conduct a two-round experiment with individual-level randomization stratified by the gender of the officers. Upon arrival, each officer receives a unique key that contains their randomized assignment for both the treatment arms and the loan bundle they will evaluate. In the second-round, loan bundles are assigned sequentially based on the first-round assignment: if an officer receives Bundle 1 in the first-round, they are assigned Bundle 2 in the second-round.

The overall randomization process is outlined in Figure 2. In the first-round, officers are randomized into one of four groups. In addition to the higher performance pay introduced in theoretical framework, I introduce a nudge-based intervention using IAT feedback and its combination with the higher performance pay. Specifically, the first group is the control group (C), where officers make approval or rejection decisions without any intervention. The second group is the IAT feedback group (T1), where officers receive feedback on their results from the Implicit Association Test (IAT) prior to making loan decisions. The feedback format, shown in Figure A3, is based on the version available at <https://implicit.harvard.edu/implicit/>. This nudge, grounded in intrinsic motivation, aims to mitigate bias by raising awareness of unconscious attitudes. The third group is the higher performance pay group (T2), where officers receive double the standard performance pay tied to ex-post loan outcomes. The incentive structure is summarized in Table A2. Specifically, the minimum compensation—if all decisions are incorrect—is 180 EGP, while the maximum—if all decisions are correct—is 640 EGP. Given that the average monthly salary of participating officers is approximately 6,000 EGP and the experiment takes about 50 minutes to complete, the change in incentive size is substantial enough to meaningfully influence behavior. This extrinsic motivation is designed to raise the opportunity cost of biased behavior and encourage greater screening effort. The fourth group is a combined intervention group (T3), integrating both IAT feedback and higher performance pay. Officers in this group first receive

feedback on their IAT results, then an explanation of the higher performance pay structure. This combined treatment aims to simultaneously leverage both intrinsic (nudge) and extrinsic (financial) motivation.

After completing the loan decisions in the first-round, officers are assigned a different bundle of loan portfolios and further randomized into one of three groups⁹ without knowing the results from the first round, as illustrated in Figure 2. This second-round randomization is stratified within each first-round treatment group, with the original performance pay structure preserved and without disclosing first-round results to officers. The first group is the AI-Control group, where officers make approval or rejection decisions without any assistance from AI. The second group is the Non-Generative AI group, where officers receive approval or rejection recommendations, along with two justifications via a custom built AI chatbot, but are not permitted to ask follow-up questions—similar to using an improved loan-screening algorithm. Officers can only passively view AI’s recommendation as shown in Figure A4. The third group is the Generative AI group, where officers receive the same recommendations and justifications as in the Non-Generative AI group but are also allowed to interact with the chatbot by asking follow-up questions (Figure A5). To encourage engagement, they are required to ask questions in at least three of the 10 cases. Section 3.1.3 details the process used to generate the chatbot and AI responses.

3.1.3 AI-assisted Decision Design

The AI chatbot used in the second round is based on GPT-4o, but I do not use real-time AI predictions for loan approval decisions. Large Language Models are stochastic, so simply changing names or asking the same question at different times can alter their responses, making it difficult to control for factors other than gender when using real-time predictions. Instead, to maintain experimental control, I pre-assign the chatbot’s approval or rejection recommendation for each loan. However, I use GPT-4o to generate two supporting reasons for each pre-assigned decision and to respond to officers’ follow-up questions. For loan officers, the chatbot was introduced as ChatGPT, but with a different interface.

The AI’s approval or rejection recommendation is primarily based on ex-post outcomes—approving loans that were repaid and rejecting those that defaulted. To observe how officers respond to both correct and incorrect AI suggestions—particularly when the AI recommends approving loans with a high chance of default—I intentionally introduce controlled errors. In each bundle of 10 loans, there are 4 ex-post defaulted loans. Among these, the chatbot recommends approval for two of them. As a result, the chatbot achieves 80% accuracy based on ex-post outcomes: it approves all 6 repaid loans, correctly rejects 2 defaulted loans, and incorrectly approves 2 defaulted loans.

⁹The second-round experimental design was modified over three cohorts: Cohort 1 (n=160) randomized officers only between the non-generative and generative AI groups; Cohort 2 (n=240) introduced an AI control group with a 1:1:1 ratio; and Cohort 3 (n=320) used a 2:1:1 ratio to balance the overall number of participants across groups.

While this high-accuracy version of AI does not reflect GPT-4o’s real-time performance, it is not far-fetched, as AI capabilities continue to improve rapidly. Current generative AI is still limited in predictive accuracy. For example, GPT-4o correctly predicted the ex-post outcomes of only 48%¹⁰ of the 50 loans used in the experiment, based on Prompt A in Appendix C. However, machine-learning tools have already achieved higher performance than humans in some settings. For example, Kleinberg et al. 2018 show that an algorithm outperforms judges in bail decisions. These findings suggest that future AI systems could plausibly attain substantially higher accuracy. This design allows an examination of how loan officers might respond to a future version of AI with substantially improved predictive ability.

Next, using Prompt B in Appendix C, I employ GPT-4o to generate two supporting justifications for each pre-assigned approval or rejection decision, based on loan application characteristics. If officers ask follow-up questions, GPT-4o responds in real-time with the temperature set to zero.

Finally, by customizing the chatbot, I ensure that it displays the pre-assigned recommendation as the AI’s decision and presents the two justifications accordingly. Officers can view the chatbot’s response without needing to input any queries. This approach facilitates the experiment’s implementation by holding AI recommendations constant across identical loan portfolios that vary only by applicant gender, and by controlling for differences arising from user inputs.

3.2 Data

3.2.1 Balance Table

The balance of baseline characteristics based on the first-round randomization is presented in Table 1. Overall, there are no significant differences across the randomized groups, with the exception of weekly working hours. The joint F-test yields a p-value of 0.404, indicating that, collectively, the groups are well balanced on observable characteristics¹¹. To clarify the variables reported in the balance table: the monthly approval rate refers to the average proportion of loans approved out of the total assigned loans; loan screening time (hours) measures the average time spent per loan on making approval or rejection decisions. Additional details about other baseline measures are provided in Appendix D.

The balance of baseline characteristics based on the second-round randomization is presented in Table A3. The joint F-test yields a p-value of 0.787, indicating that, collectively, the second-round groups are well balanced on observable characteristics.

¹⁰This does not necessarily imply that predicting loan outcomes is inherently difficult. For example, loan officers in the first-round control group achieve a 56% accuracy rate in predicting ex-post outcomes. Moreover, the machine learning model introduced in Section 4.2.4 achieves 83% accuracy using the same ex-post outcomes.

¹¹Due to a technical error, IAT scores were not collected for 11 loan officers—specifically, those in the control group and the higher performance pay group (T2), both of which did not receive IAT feedback as part of their treatment. For these officers, missing IAT scores were imputed using the median value.

3.2.2 Ex-post Repaid vs. Defaulted Loans

Loan officer’s main task was to find which one is ex-post defaulted loans, and ex-post repaid loans based on observable characteristics. Then, the ex-post repaid loans and defaulted loans are possible to distinguishable. Using logit regression, I assess whether defaulted and repaid loans differ significantly in their financial characteristics by estimating Equation 1.

$$\text{logit}(\text{Pr}(\text{Default}_i = 1)) = \alpha + \sum_{k=1}^9 \beta_k X_{ik} \quad (1)$$

where i indexes 50 loan portfolios and k indexes 9 loan characteristics, excluding categorical variables such as industry or region due to the small loan sample size, with β_k capturing the association between each characteristic and the likelihood of default. The outcome variable Default_i equals 1 if the loan defaulted and 0 if it was repaid.

In Table 2, the joint F-test yields a p-value of 0.001, indicating that defaulted loans differ significantly from repaid loans based on observable financial characteristics. Specifically, ex-post defaulted loans tend to have larger loan amounts and longer tenures, but not higher monthly borrower income. Ex-post defaulted loans also have higher late repayment history on their most recent loans, fewer employees. While not individually statistically significant, other patterns—such as a higher number of previous loans, younger borrowers, fewer employee guarantors, and a lower likelihood of being registered with the government—also suggest that ex-post defaulted loans are, on average, of lower quality than repaid loans.

3.2.3 Implicit vs. Explicit Bias

The IAT score captures the officers’ implicit bias on connecting male with career and female with family. Figure A6 presents the distribution of officer’s IAT scores, categorized into seven ranges¹². Based on IAT score range, more than half of officers have IAT score above 0.35, which stands for medium or above association between male-career and female-family.

Figure A7 presents the distribution of officer’s explicit bias showing the average scores across six explicit gender bias questions¹³, drawn from the 2017 World Values Survey (Haerpfer et al., 2022). These explicit

¹²The ranges are coded from 3 to -3, where positive numbers indicate an association of male with career and female with family, and negative numbers indicate the opposite association. Specifically, a value of 3 corresponds to an IAT score above 0.65, indicating a strong association between male-career and female-family. A value of 2 corresponds to an IAT score between 0.35 and 0.65, reflecting a moderate male-career and female-family association. A value of 1 represents an IAT score between 0.15 and 0.35, suggesting a slight male-career and female-family association. A value of 0 reflects an IAT score between -0.15 and 0.15, indicating little or no implicit association. On the negative side, a value of -1 represents an IAT score between -0.35 and -0.15, corresponding to a slight female-career and male-family association. A value of -2 corresponds to an IAT score between -0.65 and -0.35, indicating a moderate female-career and male-family association. Finally, a value of -3 corresponds to an IAT score below -0.65, indicating a strong implicit association of female-career and male-family.

¹³Each question uses a 5-point Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree), with higher values indicating greater explicit bias.

bias questions were asked at the end of the experiment after the lending decisions, so they do not capture pure measures of explicit bias. In contrast to the implicit bias measured by the IAT, the majority of officers score below 3 on average, suggesting that they do not exhibit explicit gender bias. In particular, for the business-related statement shown in Figure A8—“On the whole, men make better business executives than women do”—fewer than half of the officers expressed explicit bias, with scores of 4 or higher.

These findings reveal an implicit-explicit bias discrepancy (IED): officers exhibit implicit bias (via IAT scores) while reporting little or no explicit bias. Notably, the pattern varies by officer gender. Male officers tend to report higher explicit bias on surveys but show no significant implicit bias, whereas female officers report low explicit bias but demonstrate implicit gender bias in IAT results. This IED pattern aligns with evidence from other contexts, such as South Africa (Chisadza et al., 2025). The heterogeneous analysis by IAT score and the gender of loan officers is presented in Section 5.

4 Results

4.1 The Existence of Gender Bias in Loan Approval

Gender bias in lending decisions is measured by observing whether same loan portfolios are more likely to be approved when assigned male names compared to female names in the control group, using Equation 2:

$$Y_{ip} = \alpha + \beta Male_p + \mu_p + \delta_{LO\ gender} + \tau_{cohort} + \epsilon_{ip} \quad (2)$$

where Y_{ip} is a binary indicator equal to 1 if loan officer i approves portfolio p , and 0 otherwise. $Male_p$ is a binary variable indicating whether the portfolio is assigned a male name. μ_p denotes portfolio fixed effects, providing a within-portfolio estimate of gender gap. $\delta_{LO\ gender}$ is the gender of the loan officer fixed effect, and τ_{cohort} is the randomization cohort fixed effect accounting for 3 different cohorts in 2nd round randomization. Standard errors are clustered at loan officer level. In this specification, β represents the approval rate difference between male and female portfolios, with a positive β indicating bias against female applicants.

Column 1 of Table 3 reports the estimates of β using the overall sample. On average, the same applications are 4.4 p.p. (6.2%) more likely to be approved with male names than with female names, providing strong evidence of gender bias in loan approvals. Specifically, loans with female names are approved at a rate of 70.8%, which exceeds the 60% ex-post repayment rate, indicating over-approval. However, this may simply reflect loan officers’ usual behavior, as they reported approving 67.3% of their assigned loans on average. In contrast, loans with male names are approved at a rate of 75.2%, which is even further above the ex-post

repayment rate.

4.2 The Nature of Gender Bias: Generosity Toward Men

The nature of gender bias can manifest as generosity toward men, harshness toward women, or a combination of both. Bias is characterized as generosity toward men when the gender gap is more pronounced in low-quality loans, and as harshness toward women when the gap is more concentrated in high-quality loans. To distinguish between high- and low-quality loans, I use three measures: (1) ex-post repayment outcomes, (2) past late repayment history, and (3) credit scores.

4.2.1 Ex-post Outcome

The ex-post loan outcome—whether repaid or defaulted—offers indication of quality, with defaulted loans more likely to have been low-quality and repaid loans more likely to have been high-quality on average. As discussed earlier with Table 2, ex-post defaulted loans exhibit lower-quality characteristics on average compared to ex-post repaid loans. Using Equation 2, I conduct separate subgroup analyses for ex-post defaulted and ex-post repaid loans. Column (2) of Table 3 shows that the same loan portfolio is approved 9.6 p.p. (14.5%) more often when assigned a male name compared to a female name, suggesting generosity toward men. In contrast, among ex-post repaid loans (Column (3)), there is no longer a statistically significant gender gap in approval rates, indicating that harshness toward women is not evident.

However, ex-post outcomes have limitations due to their stochastic nature: repaid loans are not necessarily ex-ante high-quality, nor are defaulted loans necessarily ex-ante low-quality. Moreover, loan officers could not observe ex-post outcomes during the experiment. To address this limitation, I provide additional evidence of generosity toward men using two alternative, observable measures of quality: past late repayment history and credit scores.

4.2.2 Late Repayment History

Out of the 50 loans, 11 have a history of late repayment on their most recent loans, meaning the borrower was at least one day late on their most recent loan repayment. Unlike the ex-post loan outcome—which is not observable to officers during the experiment—late repayment history is observable to officers and serves as a clear signal of default risk. Using Equation 2, I conduct subgroup analyses separately for loans with late repayment history and those without it, including first-time loan applicants.

Column (1) of Table 4-1 shows that, among loans with a history of late repayment, the same portfolio is approved 11.1 p.p. (21.6%) more often when assigned a male name. In contrast, there is no statistically

significant gender difference in approval rates for loans without a late repayment history (column (1)). This provides additional evidence that the nature of gender bias in lending is driven by generosity toward men, rather than harshness toward women.

4.2.3 Credit Score

Credit scores, generated by I-Score from the Egyptian Credit Bureau, serve as one of the key indicators of a borrower's ex-ante risk of default. In this experiment, loan officers were provided with a one-page explanation detailing the tiers of credit scores: a score of 751 or above is classified as excellent; 701 to 750 as very good; 626 to 700 as satisfactory; 521 to 625 as unsatisfactory; 520 or below as bad; and "missing" indicates that a credit score does not exist. Using Equation 2, I conduct subgroup analyses based on these credit score tiers. First, I examine 16 loans with credit scores of 700 or below, which fall into the satisfactory or lower categories and are considered relatively low-quality. If gender bias appears more strongly in this group, it would suggest generosity toward men. Second, I analyze 18 loans with credit scores of 701 or above, which are relatively safer, where a stronger gender gap in favor of women would indicate harshness toward men. Finally, I analyze 16 loans without an available credit score.

Column (1) of Table 4-2 shows that, among loans with credit scores of 700 or below, the same portfolio is approved 9.7 p.p. (16.8%) more often when assigned a male name, which provides evidence of generosity toward men. In contrast, Column (2) shows that among loans with credit scores of 701 or above, there is no statistically significant gender gap in approval rates. Column (3) reveals that even among loans without a credit score, there is no statistically significant gender difference in approval. This result is unexpected, given that the absence of a credit score could signal lower quality.

However, the lack of gender bias in loans without credit scores is likely to be due to endogeneity in the sample of approved loans. Since the experiment includes only loans that were approved in the past, those without a credit score tend to have smaller loan amounts relative to the borrower's income. This pattern suggests that loan officers may have approved smaller loans to reduce risk, or borrowers may have applied for smaller loans to improve their chances of approval. In Table 4-2, the female approval rate for loans without credit scores is 76%, which is similar to the approval rate for loans with very good credit scores. Moreover, the repayment-to-income ratio for these loans is only 0.287, lower than that of loans in the lower credit score group. This indicates that loans without credit scores were, on average, relatively high-quality loans.

Next, I analyze whether there is a statistically significant monotonic relationship between credit score and loan approval during the experiment. Figure 3 shows that female loan approval rates have a statistically significant monotonic relationship with credit scores, as indicated by an upward-sloping line. In contrast, the relationship is not statistically significant for male portfolios. Moreover, the difference in slopes between

male and female portfolios is statistically significant at the 10% level. The insignificant slope for male portfolios and the steeper, significant slope for female portfolios suggest a pattern of generosity toward men. Specifically, male portfolios receive relatively high approval rates even at lower credit scores, resulting in a flatter approval trend across the credit score range of 680 to 750. This implies that men are more likely to be approved for loans despite having lower credit scores.

Taken together, the evidence from credit scores supports the previous findings using ex-post loan outcomes and late repayment history. Across all three measures of loan quality, the gender gap in approval rates is concentrated in low-quality loans. This consistent pattern reinforces the interpretation that the nature of gender bias in loan approval is driven by generosity toward men, rather than harshness toward women.

4.2.4 Robustness Check: ML-based Benchmarking of quality

As a robustness check to benchmark ex-ante repayment probability, I use a Gradient Boosting Machine (GBM) model¹⁴ to predict repayment probabilities for the 50 experimental loans. The model is trained¹⁵ on a dataset of 561 loans from the same MFI, using 9 baseline features¹⁶. I exclude categorical variables such as industry and region due to the limited sample size. In addition, as the credit scoring system was recently updated, the original application-time credit scores are no longer accessible and thus not included in the training data. Consequently, the ML-based predicted repayment probabilities have limitations, as they were trained on information that differs from what loan officers originally observed. The most predictive features, in descending order, are loan tenure, monthly repayment-to-income ratio, days of past late repayment, number of previous loans, and other borrower characteristics.

The model performs well, with an average AUC of 0.88, accuracy of 0.83¹⁷. For each of the 50 experimental loans, I assign the average predicted repayment probability out of 1,000 runs. Since the ML-predicted repayment probabilities differ from those inferred by loan officers¹⁸, defining a single approval threshold is challenging. I therefore use two thresholds: $p = 0.50$ as a neutral benchmark, and $p = 0.54$ as the point at which loans are classified into the bottom 40% and top 60%¹⁹. Based on these thresholds, I define the

¹⁴The GBM model is chosen for its superior Area Under the ROC Curve (AUC) performance compared to Random Forest, Support Vector Machines, neural networks, XGBoost, and CatBoost on this dataset.

¹⁵To address class imbalance, I apply SMOTE and perform cross-validation with 10 folds and 10 repeats. Hyperparameters are tuned over a grid: the number of trees ranges from 300 to 1000 (in increments of 100); tree depth is set at 3, 5, or 7; the learning rate is 0.005; and the minimum number of observations per terminal node is 10.

¹⁶Monthly repayment-to-income ratio, loan tenure (months), days of past late repayment, number of previous loans, age of borrower, registered business (binary), first-time loan (binary), number of employees, number of employee guarantors.

¹⁷Using the 0.6 probability threshold, the model achieves an F1 score of 0.85, precision of 0.86, and recall of 0.84—each with a standard deviation of approximately 0.01 across 1,000 iterations.

¹⁸The ML model is trained using only approved loans, which likely excludes very low-quality applications. As a result, the model may overestimate the probability of repayment. In addition, credit score, industry, and geographic variables are excluded from model training due to the limited size of the training dataset and data availability.

¹⁹The ex-post default rate is 40%, and loan officers are informed of this, although the exact rate varies across individuals. Accordingly, the bottom 40% of loans, based on predicted repayment probabilities from the machine-learning model, can be considered low-quality loans.

gender gap below the threshold as reflecting generosity toward men (approving more low-quality loans with male names), and the gender gap above the threshold as reflecting harshness toward women (rejecting more high-quality loans with female names).

Table A4 presents the regression results using Equation 2. Column (1) shows that among low-quality loans (predicted repayment probability below 0.5), the gender gap is 6.8 p.p. (10.4%) and statistically significant. In contrast, Column (2), which includes high-quality loans (probability 0.5 or above), shows a smaller and statistically insignificant gender gap of 3.1 p.p. (4.2%). Similarly, using a threshold of 0.54, Column (3) shows that the gender gap among low-quality loans is 6.1 p.p. (9.1%) and statistically significant, while Column (4) shows a gap of only 3.3 p.p. (4.5%) among high-quality loans, which is not statistically significant. Although I cannot statistically reject the null hypothesis that the gender gap in low-quality loans equals that in high-quality loans, the pattern suggests that the bias is consistent with generosity toward men.

4.3 Debiasing Treatment Effect

4.3.1 Change in Gender Gap in Loan Approval

The effects of the three debiasing treatments on loan approval are analyzed using Equation 3.

$$Y_{ip} = \alpha + \sum_{k=1}^3 \beta_k T_{ik} Male_p + \sum_{k=1}^3 \gamma_k T_{ik} + \delta Male_p + \mu_p + \delta_{LO \text{ gender}} + \tau_{\text{cohort}} + \epsilon_{ip} \quad (3)$$

where Y_{ip} is a binary indicator equal to 1 if loan officer i approves portfolio p , and 0 otherwise. T_{ik} is a binary variable that equals 1 if loan officer i belongs to treatment group k , and 0 for all other groups. The k can take values 1, 2, or 3, corresponding to the three treatment groups. All other definitions remain the same as in Equation 2. The interaction coefficient β_k between the treatment dummy T_{ik} and the male applicant dummy $Male_p$ captures the change in the gender gap in loan approval in treatment group k , relative to the gender gap in the control group, captured by δ .

Column (1) of Table 5 presents the estimated coefficients β_k from Equation 3. The results show that the IAT feedback alone does not statistically significantly reduce the gender gap in loan approval. However, higher performance pay and the combined treatment of higher performance pay with IAT feedback reduce the gender gap by 6.6 p.p. and 4.8 p.p., respectively, effectively closing it.

Columns (2) and (3) present subgroup analyses of Column (1) by the Gender–Career IAT score. Column (2) includes officers biased against women (IAT score ≥ 0.35), and Column (3) includes the rest. The debiasing effect of IAT feedback should be analyzed primarily among officers whose IAT scores indicate bias against women. The feedback is expected to have a stronger impact among these officers, as it explicitly

highlights their implicit association between men and careers. For officers without such bias, the feedback simply confirms neutrality (“you are not associating men with career and women with family”) and is therefore unlikely to induce behavioral change; any observed effect would likely reflect experimental noise.

However, contrary to this expectation, although statistically insignificant, the IAT feedback reduces the gender gap by only 2.6 p.p. among officers biased against women, compared with 4.0 p.p. among others. This suggests that officers who are already aware of their bias are unlikely to change their behavior merely by being told that they associate men with careers and women with family, without additional monetary incentives. On the other hand, the effects of higher performance pay and the combined treatment are suggestively greater for officers biased against women.

4.3.2 Change in Screening Performance and Screening Time

Next, I analyze whether the debiasing treatments enhance loan officers’ screening performance and alter their screening time, thereby assessing changes in credit allocation efficiency using Equation 4.

$$Y_{ip} = \alpha + \sum_{k=1}^3 \gamma_k T_{ik} + \mu_p + \delta_{LO\ gender} + \tau_{cohort} + \epsilon_{ip} \quad (4)$$

where Y_{ip} denotes the outcome variable. In Columns (1) and (2), it is a binary accuracy indicator for male-name and female-name loans, equal to 1 if the officer approves an ex-post repaid loan or rejects an ex-post defaulted loan, and 0 if the officer approves an ex-post defaulted loan or rejects an ex-post repaid loan. In Column (3), Y_{ip} represents the average screening time per loan²⁰. All other variable definitions remain the same as in Equation 3.

Columns (1) and (2) of Table 6 present the change in screening accuracy in each treatment arm relative to the control group for male-name and female-name loans, respectively. Higher performance pay and the combined intervention increase the screening accuracy for male-name loans by 3.6 p.p. (6.6%) and 4.2 p.p. (7.7%), respectively, while the IAT feedback alone has no statistically significant effect. In contrast, there is no statistically significant change in the screening performance for female-name loans across any treatment arm²¹. Column (3) reports the change in average screening time per loan across treatments. Although statistically insignificant, officers receiving IAT feedback reduce their screening time by 0.03 minutes (3.1%) per loan, whereas officers with higher performance pay increase it by 0.12 minutes (11.3%) per loan. Officers under the combined intervention increase their screening time slightly by 0.01 minutes (0.9%) per loan.

²⁰The increase in screening time reflects the overall change in officer effort across all loan applications. Screening time data broken down by the applicant’s gender was not collected during the experiment.

²¹Screening performance measured using the ML model is presented in Table A5, showing no statistically significant change across treatments. This may be because the ML-based repayment probability differs from the model used by loan officers, as certain information—such as credit scores—is missing. In addition, the control group’s accuracy measured by the ML model does not differ between male-name and female-name loans, which is counterintuitive.

Interpreting the treatment effects of the combined intervention, combining IAT feedback with higher performance pay does not produce simple additive effects but instead reveals a more complex interplay between the two interventions. One possibility is that, with the clearer guidance provided by IAT feedback, loan officers used their screening time more efficiently. Consistent with this, screening time slightly decreased under IAT feedback (though not significantly), perhaps because officers had clearer direction. By contrast, under higher performance pay alone, officers responded by increasing screening time. In the combined intervention, officers spent an intermediate amount of time relative to each treatment, while male-name screening performance improved in a way similar to higher performance pay. This pattern suggests that officers may have used their time more effectively when IAT feedback was paired with performance pay.

These results are also consistent with the testable predictions from Section 2: Testable Prediction 1, which states that “Loan officers with the higher performance pay treatment are expected to spend more time on loan screening relative to officers in the control group,” and Testable Prediction 2, which posits that “If the pro-male gender gap in loan approval is driven by generosity toward men, higher performance pay is expected to increase the approval threshold for male applicants”.

4.4 AI-Assisted Lending

Improving screening accuracy remains a critical challenge in low-income countries, where default rates are high but loan officers often have limited skills and training resources. The IAT feedback, higher performance pay, and combined interventions help reduce gender bias but do not substantially improve screening performance. These interventions work only within the limits of existing human ability and cannot fundamentally raise screening capacity. Motivated by these limitations, I next turn to AI-assisted decision-making. Using the second-round loan decisions, I test whether officers assisted by a high-performing AI tool make better screening choices than those in the control group without AI assistance, and whether they use AI recommendations neutrally by applicant gender—that is, whether they follow AI guidance differently for male- versus female-name applications.

4.4.1 AI for Better Loan Screening

I first test whether loan officers assisted by AI make better screening decisions compared to those in the control group without AI using Equation 5.

$$Y_{ip} = \alpha + \sum_{k=1}^2 \gamma_k T_{ik} + \delta Male_p + \mu_p + \delta_{LO \text{ gender}} + \tau_{cohort} + \epsilon_{ip} \quad (5)$$

where Y_{ip} is binary screening performance variable that equals 1 if loan officer i makes the correct decision about portfolio p , and 0 if they make an incorrect decision. T_{iK} is a binary variable that equals 1 if loan officer i is randomized into Treatment k , but 0 for otherwise. In the second round, there are two treatment groups: the generative-AI group ($k = 1$), in which officers can actively interact with the AI (similar to ChatGPT), and the non-generative-AI group ($k = 2$), in which officers can only passively view the AI's suggestions, as in a traditional algorithmic system. All other variables are defined as in Equation 4.

Column (1) of Table 7 presents the screening performance²² gains of loan officers using AI, combining both generative and non-generative AI users. The results indicate that officers using AI achieve a 7.8 p.p. (14.1%) improvement in screening performance compared to those in the control group without AI assistance. Column (2) disaggregates the results by AI type (γ_k) and shows that both generative and non-generative AI users exhibit improved screening performance relative to the control group. However, there is no additional performance gain from using generative AI, despite the fact that generative AI users spent 45.3% more time on loan screening

As a robustness check, I use an alternative measure of screening performance based on the repayment probabilities predicted by the ML model from Section 4.2.4. Table A6 shows that, loan officers assisted by either generative or non-generative AI exhibit better screening performance than those without AI assistance.

Overall, the results show that AI-assisted decision-making (human + AI) may perform worse than automated AI decisions. Moreover, there is suggestive evidence that officers using generative AI may underperform relative to those using non-generative AI. In the ex-post outcome measure reported in Table 7, the AI model achieves 80% accuracy. In comparison, officers using generative AI achieve 61.9% accuracy, while those using non-generative AI reach 64.2%. These findings are consistent with Agarwal et al. (2025), showing that when AI predictions are highly confident, human-AI collaboration can underperform relative to automated AI decisions due to human overconfidence and suboptimal overrides.

4.4.2 Generosity Toward Men in Following AI Suggestions

To explore why generative AI users may exhibit lower screening performance gains compared to non-generative AI users, I analyze their rate of following AI recommendations across three types of AI recommendations: (1) approval of ex-post repaid loans, (2) approval of ex-post defaulted loans, and (3) rejection of ex-post defaulted loans. To conduct this analysis, I estimate Equation 6, which tests whether loan officers in the generative AI group are more likely to follow AI recommendations for male-name portfolios compared to officers in the non-generative AI group.

²²Screening performance is defined as a binary variable based on ex-post loan outcomes: it equals 1 if the officer approves a loan that ex-post repaid or rejects a loan that ex-post defaulted, and 0 otherwise.

$$Y_{ip} = \alpha + \beta T_i Male_p + \gamma T_i + \theta Male_p + \mu_p + \delta_{LO\ gender} + \tau_{cohort} + \epsilon_{ip} \quad (6)$$

where Y_{ip} is a binary which is equal to 1 if loan officer i follows AI's decision on portfolio p , and 0 otherwise. T_i is a binary variable that equals 1 if loan officer i is randomized into generative AI group, but 0 for non-generative AI group. All other variables are defined as in Equation 6.

Table 8 presents the estimated coefficients β , γ , and θ , corresponding to the analysis described in Equation 6. Column (1) shows that for both generative and non-generative AI users, there is no significant difference in following AI's rejection recommendations for ex-post defaulted loans, indicating no evidence of bias in rejecting low-quality loans across portfolio genders. Column (2) focuses on cases where AI incorrectly recommends approval for ex-post defaulted loans. The results show that non-generative AI users do not follow these recommendations differently by gender of applicants. However, generative AI users follow these incorrect approval recommendations 11.3 p.p. more often for male-name portfolios compared to non-generative AI users. This suggests that generative AI users exhibit generosity toward men when using generative AI, where interactive communication is possible. Column (3) examines follow rates when AI correctly recommends approval for ex-post repaid loans. In this case, there is no statistically significant difference in follow rates by gender for either generative or non-generative AI users, indicating that officers follow accurate approval recommendations similarly across applicant genders.

As a robustness check, I examine whether loan officers using generative AI are more likely to follow AI approval recommendations for low-quality loans when evaluating male-name applications, using ML-predicted repayment probabilities from Section 4.2.4 to define quality. Table A7 presents the estimated coefficients β , γ , and θ from the analysis described in Equation 6. Column (1) reports results when low-quality loans are defined as those with ML-predicted repayment probabilities below 0.4. In this case, generative AI users follow AI approval recommendations 10.1 p.p. more than non-generative AI users. In Column (2), using a threshold of 0.5 to define low-quality loans, the difference is 8.8 p.p. In Column (3), using a 0.6 threshold, the difference narrows to 5.9 p.p. and is no longer statistically significant. It suggests a pattern of generosity toward male applicants, with loan officers using generative AI more likely to follow approval recommendations for low-quality loans when the applicant has a male name.

4.4.3 Generosity Toward Men in Interactions with AI

To better understand the mechanism behind the selective follow rate of AI's recommendation by gender of applicants, I analyze the communication between loan officers and the generative AI chatbot during the experiment. Table A8 presents summary statistics on the questions asked to the chatbot. While officers

were instructed to ask questions to the chatbot for at least 3 out of the 10 portfolios, they asked questions for only 2.3 portfolios on average. Moreover, 93% of interactions consisted of a single question, with only 7% involving further follow-up with the chatbot. The likelihood of asking a question varied by the AI’s recommendation: questions were asked for 20% of AI-approved ex-post repaid loans, 19% of AI-approved ex-post defaulted loans, and 37% of AI-rejected ex-post defaulted loans. In Table A9, I categorize the types of questions posed to the chatbot into three groups: (1) information-seeking or neutral, (2) weak disagreement, and (3) strong disagreement. The distribution shows that 40% of questions were information-seeking or neutral, 50% expressed weak disagreement, and 10% involved strong disagreement with the AI’s recommendation.

Table A10 presents whether the same portfolio is more likely to prompt a question when it has a female versus male name using Equation 2 within generative AI users only. The outcome is a binary variable equal to 1 if the portfolio is asked any questions by officers, and 0 otherwise. Interestingly, when the AI recommends approval of ex-post defaulted loans, female-name portfolios are questioned 8% more often than male-name portfolios. Next, Table A11 analyzes the intensity of disagreement with the AI, based on the gender of the portfolio name using Equation 2 within generative AI users only. The outcome variable is an intensity scale ranging from 0 to 3, where 0 indicates no question, 1 denotes an information-seeking or neutral question, 2 a weak disagreement, and 3 a strong disagreement. Column (3) shows that, when the AI recommends approval of ex-post defaulted loans, female-name portfolios receive questions with 0.15 higher intensity, on average, than male-name portfolios.

Taken together, these findings suggest that loan officers with pre-existing gender bias stemming from generosity toward men interact with generative AI in ways that may align with their bias. This implies that communication with AI does not necessarily mitigate bias; rather, its effectiveness depends not only on the fairness of the AI system but also on the behaviors of its users.

5 Heterogeneous Effects

5.1 Gender Gap in Approval

I find that loan officers are 4.4 p.p. (6.2%) more likely to approve the same portfolio when it is assigned a male name. In this section, I examine the heterogeneity of this gender bias by estimating Equation 2 and exploring variation separately by each of four loan officer characteristics: gender, IAT score, years of experience, and education level.

Table A12 presents the results. Columns (1) and (2) show that both male and female loan officers exhibit

a similar degree of gender bias in approval decisions, indicating that the officer’s gender does not significantly affect the extent of bias. In Columns (3) and (4), I divide officers based on their IAT scores. Officers with scores equal to or above 0.35 are classified as having moderate or stronger implicit associations linking men with careers and women with family roles, following prior studies (Greenwald et al., 1998, 2003; Lane et al., 2007). These officers exhibit a 6.9% (4.8 p.p.) gender gap in loan approval, compared to 4.2% (3.0 p.p.) among officers with lower IAT scores. However, the difference between these groups is not statistically significant. Lastly, I examine heterogeneity by experience, splitting officers at the median level of experience. Columns (5) and (6) show that officers with below-median experience exhibit a 7.3 p.p. gender gap, while those with above-median experience show only a 1.8 p.p. gap. Although the difference is not statistically significant, the pattern suggests that bias may be stronger among less experienced officers, consistent with prior evidence (Alibhai et al., 2019; Brock and De Haas, 2023). Lastly, Columns (7) and (8) show that officers with below-college education exhibit an 8.3 p.p. gender gap, whereas those with a college or higher education show only a 2.5 p.p. gap. While this difference is also not statistically significant, it suggests that bias may be more pronounced among officers with lower education levels.

5.2 Debiasing Treatment Effects

Using the same officer characteristics—gender, IAT score, and years of experience—I examine heterogeneity in the effectiveness of the three debiasing treatments by estimating Equation 3 separately for each characteristic. The results are presented in Table A13.

By the gender of loan officers and their IAT scores, the results suggest that female officers may respond more strongly to the IAT feedback, while officers with higher IAT scores (i.e., those with stronger implicit bias) appear to respond more to the higher performance pay. However, these differences across subgroups are not statistically significant. Columns (3) and (4) show that the IAT feedback reduces the gender gap by only 2.6 p.p. among biased officers, which is smaller than the effect among less biased officers. This pattern suggests that officers with stronger implicit bias may already be aware of their attitudes, and therefore IAT feedback, which is intended to reveal unconscious bias, offers little additional information or behavioral adjustment. In Columns (5) and (6), the debiasing effect of higher performance pay is larger for less experienced officers. This result indicates that less experienced officers may not only exhibit greater gender bias but also respond more strongly to monetary incentives. Lastly, Columns (7) and (8) suggest that the debiasing effect of higher performance pay is larger among officers with lower education levels.

5.3 Loan Screening Performance Gain from AI Use

Loan officers with AI assistance exhibit improved screening performance compared to those without it. Using Equation 5, I analyze heterogeneity in the screening performance gains from AI assistance by officer characteristics. Specifically, I estimated it separately by four characteristics: gender, IAT score, years of experience as a loan officer, and prior experience using AI. The results are presented in Table A14.

In Columns (1) and (2), both male and female officers show similar levels of improvement in screening performance, regardless of whether they used generative or non-generative AI. Columns (3) and (4) show no statistically significant difference in the screening performance gains from using AI by implicit gender bias, as measured by IAT scores. Columns (5) and (6) reveal an interesting pattern regarding experience of loan officers. Inexperienced officers show greater screening performance gains from using generative AI compared to experienced officers. However, there is no statistically significant difference in the effect of non-generative AI by officers' experience level. It suggests that experienced officers rely more on their own judgment rather than AI input, resulting in smaller accuracy improvements.

The bottom panel of Table A14 presents heterogeneity by loan officers' education level and prior AI experience. However, there is no statistically significant difference in screening performance gains across education levels or prior AI experience.

5.4 Selective Use of AI by Applicant Gender

Loan officers selectively follow AI recommendations by the gender of the applicant, specifically when the generative AI recommends approval for ex-post defaulted loans. Using Equation 6, I further analyze heterogeneity in selective adherence by officer characteristics—gender, IAT score, years of experience as a loan officer, education level, and prior experience using AI. Results are presented in Table A15.

In Columns (1) and (2), both male and female officers exhibit similar patterns: they are more likely to follow AI recommendations to approve ex-post defaulted loans when using generative AI than when using non-generative AI. Similarly, Columns (3) to (6) show no significant heterogeneity based on the implicit gender bias as measured by the IAT score and experience level.

However, in the lower panel of Table A15 show heterogeneity by officers' education level and previous AI experience. Column (1) shows that officers with college or above education level follow AI's approval recommendations significantly more for male-named applications when using generative AI compared to non-generative AI—by a margin of 19.9 p.p., whereas no such pattern is observed among officers with lower education levels. Similarly, Column (3) shows that officers with prior AI experience follow AI's approval recommendations significantly more for male-named portfolios under generative AI, by 23.9 p.p., while no

comparable effect appears among those without prior AI experience. These results suggest that officers with higher education or prior AI experience are more likely to use AI selectively. Notably, officers with higher education do not exhibit gender bias in the first round but appear to use AI in a gender-biased manner. A plausible interpretation is that when AI offers a rationale that seems to legitimize biased behavior, these officers become more willing to act on that bias.

6 Policy Implications and Conclusion

In summary, the findings highlight the existence of gender bias²³ in small business lending, with the nature of the bias driven by generosity toward men, rather than harshness toward women. Higher performance pay helps close this gender gap, and AI assistance appears promising, though user bias must be addressed to ensure gender-neutral use of AI. Although the quantitative magnitudes may differ in other contexts, the framework of generosity and harshness, the scalable performance-pay solution, and the importance of accounting for users’ biases in AI use have external validity beyond Egypt and the lending context.

According to the Central Bank of Egypt²⁴, as of end-June 2023, the total outstanding loan portfolio of the banking sector stood at EGP 4,798.9 billion, with a non-performing loan (NPL) ratio of 3.3%²⁵. A back-of-the-envelope calculation²⁶ suggests that 6.9% of NPLs are attributable to generosity toward men. The resulting financial loss is approximately EGP 3.6 billion (USD 72.9 million²⁷) per year. In other words, closing the gender bias among loan officers could potentially increase bank profits by up to this amount—resources that could be used to lower interest rates or expand credit access. On the other hand, the estimated increase in labor costs from higher performance pay is about EGP 429.1 million annually²⁸.

²³The experiment is designed to capture taste-based bias (Becker, 1957), although both taste-based and statistical discrimination are likely to operate jointly. To mitigate inaccurate statistical discrimination (Bohren et al., 2023), the experiment focuses on industries with relatively gender-balanced borrower representation (Agriculture, supermarkets, grocery stores, clothing shops, fish sale shops, poultry, and livestock) and informs officers that default rates do not differ by applicant gender in the sample. There is suggestive evidence that taste-based bias is at play. In an endline question that primarily reflects statistical discrimination—“On the whole, men make better business executives than women do”—the experimental gender gap in loan approval does not differ between officers who agree with the statement and those who do not. Moreover, higher performance pay eliminates the gender gap while increasing screening time by only 11%, which is also consistent with taste-based bias alongside statistical discrimination. In contrast, the finding that the gender gap decreases with officers’ experience suggests the statistical discrimination channel.

²⁴Central Bank of Egypt, *Annual Report 2022/2023*, <https://www.cbe.org.eg/en/economic-research/economic-reports/annual-report>

²⁵The current NPL rate is relatively low compared to historical peaks—10.9% in 2011 and 26.5% in 2005 (Federal Reserve Bank of St. Louis, <https://fred.stlouisfed.org/series/DDS1o2EGA156NWDB>)

²⁶Out of ex-post defaulted loans, 66.1% of female-named loans were approved compared to 75.8% of male-named loans. Without generosity toward men, male-named loans would also have been approved at 66.1%, the same rate as female-named loans. In that case, ex-post defaulted loans overall would have been approved at 66.1%, rather than the observed 71.0% during the experiment. This implies that generosity toward men accounts for $\frac{71.0 - 66.1}{71.0} = 6.9\%$. Applying this margin to the stock of loans (EGP 4,798.9 billion $\times 0.033 \times 0.069$) yields an estimated EGP 10.9 billion. Since mortgage loans—typically with longer maturities—constitute only 9.8% of retail loans (Central Bank of Egypt, Financial Stability Report, March 2024, <https://www.cbe.org.eg/en/financial-stability/macprudential/financial-stability-reports>), assuming an average loan tenure of three years implies an annualized loss of roughly EGP 3.6 billion (USD 72.9 million).

²⁷At an exchange rate of 1 USD = 50 EGP.

²⁸The higher performance pay raised participation rewards by 2.1% in the experiment. The average annual in-

Taken together, these findings imply that increasing performance pay could raise annual net profits by roughly EGP 3.2 billion (USD 64.3 million). The negative consequences of generosity toward men may be particularly pronounced during banking crises, when higher default rates amplify systemic risk and exacerbate economic shocks, especially in low-income countries. Beyond reducing immediate bank losses, the benefits of addressing generosity bias are likely to extend further, as doing so can mitigate capital misallocation and enhance overall productivity. Hence, the total gains from reducing generosity toward men in lending likely exceed the estimated EGP 3.2 billion (USD 64.3 million). Expanding performance pay tied to repayment outcomes (e.g., monthly) is a scalable solution that can be implemented in lending institutions worldwide.

This experiment, however, has several limitations. First, although the loan officers’ decisions were incentivized, they did not have real-world consequences for actual clients in the hypothetical setting, limiting how closely the observed behavior reflects true decision-making—as discussed in other incentivized experimental studies (Kessler et al., 2019). Second, the experiment draws only on previously approved loans with intermediate or missing credit scores and a small loan sample size, thus limiting the external validity of the findings. Third, although officers were encouraged to interact with the generative AI chatbot, 93% of them engaged in only a single round of interaction, suggesting that the full potential of generative AI was not realized. Finally, the AI chatbot in the experiment was designed to have 80% accuracy, which may exceed the performance of currently deployed systems, thus potentially overstating the effectiveness of AI.

Despite these limitations, the findings have important implications for MFIs and governments. To curb the over-approval of low-quality male applicants, expanding performance-based incentives needs to be considered, particularly for less experienced loan officers, who are more responsive to the higher performance pay. On the other hand, the IAT feedback alone may be insufficient to close the gender gap. Furthermore, while AI tools may improve decision quality, especially for less experienced officers, their implementation requires care. To fully realize the gains from AI, loan officers must receive adequate training on how to use the tools and gender-sensitivity education to avoid the biased use of AI.

This study also opens several avenues for future research. First, beyond the supply-side aspects of the gender gap in lending, demand-side factors must also be considered, such as difficulties in the loan application process or behavioral patterns that may contribute to the observed gap. Second, a follow-up study could use ex-ante clearly high- and low-quality loan applications—including previously rejected ones—identified by a machine learning model. It would also be valuable to examine whether loan officers with higher performance

come of workers in the financial intermediation sector is estimated at EGP 149,928 (<https://censusinfo.capmas.gov.eg/metadata-ar-v4.2/index.php/catalog/1920/download/6926>, Central Agency for Public Mobilization and Statistics of Egypt, *Annual Bulletin of Employment, Wages and Working Hours Statistics*, 2023), and the total number of bank employees in 2022 was 136,273 (<https://www.cbe.org.eg/-/media/project/cbe/listing/monthly-statistical-bulletin/bulletin/feb/monthly-statistical-bulletin-311.pdf>, Central Bank of Egypt, *Monthly Statistical Bulletin* (February 2023)). The implied increase in labor costs is $136,273 \times 149,928 \times 0.021 \approx$ EGP 429.1 million.

pay continue to exhibit debiased behavior over time and what happens when their performance pay returns to the original level. Moreover, a clearer experimental setting that can disentangle taste-based bias from statistical discrimination would yield further insights. Third, given the rapid improvement of AI, a future study could explore how officers use real-time outputs from generative AI, building on this experiment's use of a controlled system with pre-set responses. Lastly, future research can further examine whether AI-assisted lending improves access to finance, particularly for historically underserved groups.

References

- Agan, A. and S. Starr (2018). Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics* 133(1), 191–235.
- Agan, A. Y., D. Davenport, J. Ludwig, and S. Mullainathan (2023). Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Technical report, National Bureau of Economic Research.
- Agarwal, N., A. Moehring, and A. Wolitzky (2025). Designing human-ai collaboration: A sufficient-statistic approach. Technical report, National Bureau of Economic Research.
- Agarwal, S., B. C. Morais, A. Seru, and K. Shue (2024). Noisy experts? discretion in regulation. Technical report, National Bureau of Economic Research.
- Agarwal, S., C. B. Muckley, and P. Neelakantan (2023). Countering racial discrimination in algorithmic lending: A case for model-agnostic interpretation methods. *Economics letters* 226, 111117.
- Agier, I. and A. Szafarz (2013). Microfinance and gender: Is there a glass ceiling on loan size? *World development* 42, 165–181.
- Aigner, D. J. and G. G. Cain (1977). Statistical theories of discrimination in labor markets. *Ilr Review* 30(2), 175–187.
- Alesina, A., M. Carlana, E. La Ferrara, and P. Pinotti (2024). Revealing stereotypes: Evidence from immigrants in schools. *American Economic Review* 114(7), 1916–1948.
- Alesina, A. F., F. Lotti, and P. E. Mistrulli (2013). Do women pay more for credit? evidence from italy. *Journal of the European Economic Association* 11(suppl_1), 45–66.
- Alibhai, S., A. A. Donald, M. Goldstein, A. A. Oguz, A. Pankov, and F. Strobbe (2019). Gender bias in sme lending: Experimental evidence from turkey. *World bank policy research working paper* (9100).
- Altonji, J. G. and R. M. Blank (1999). Race and gender in the labor market. *Handbook of labor economics* 3, 3143–3259.
- Amaral, S., G. Borker, N. Fiala, N. Prakash, H. Rainer, and M. M. Sviatschi (2021). Debiasing law enforcement officers: Evidence from an expressive arts intervention in india. Technical report, Working Paper.

- Andreeva, G. and A. Matuszyk (2019). The law of equal opportunities or unintended consequences?: The effect of unisex risk assessment in consumer credit. *Journal of the Royal Statistical Society Series A: Statistics in Society* 182(4), 1287–1311.
- Angelova, V., W. S. Dobbie, and C. Yang (2023). Algorithmic recommendations and human discretion. Technical report, National Bureau of Economic Research.
- Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2022). Machine bias. In *Ethics of data and analytics*, pp. 254–264. Auerbach Publications.
- Arnold, D., W. Dobbie, and P. Hull (2020). Measuring racial discrimination in bail decisions. Technical report, National Bureau of Economic Research.
- Arnold, D., W. Dobbie, and P. Hull (2022). Measuring racial discrimination in bail decisions. *American Economic Review* 112(9), 2992–3038.
- Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *The Quarterly Journal of Economics* 133(4), 1885–1932.
- Ayalew, S., S. Manian, and K. Sheth (2023). Discrimination and access to capital: Experimental evidence from ethiopia.
- Baert, S. (2018). *Hiring discrimination: An overview of (almost) all correspondence experiments since 2005*. Springer.
- Bao, H. X. (2024). Is hiding my first name enough? using behavioural interventions to mitigate racial and gender discrimination in the rental housing market. *Urban Studies* 61(11), 2156–2175.
- Barasinska, N. and D. Schäfer (2010). Are women more credit-constrained than men? evidence from a rising credit market. Technical report, FINES Working Paper.
- Bartlett, R., A. Morse, R. Stanton, and N. Wallace (2022). Consumer-lending discrimination in the fintech era. *Journal of Financial Economics* 143(1), 30–56.
- Bartős, V., S. Castro, K. Czura, and T. Opitz (2024). Gendered access to finance: The roles of team formation, idea quality, and implementation constraints in business evaluations. Technical report, CESifo Working Paper.
- Bayer, P., F. Ferreira, and S. L. Ross (2018). What drives racial and ethnic differences in high-cost mortgages? the role of high-risk lenders. *The Review of Financial Studies* 31(1), 175–205.

- Beaman, L., R. Chattopadhyay, E. Duflo, R. Pande, and P. Topalova (2009). Powerful women: does exposure reduce bias? *The Quarterly journal of economics* 124(4), 1497–1540.
- Beck, T., P. Behr, and A. Madestam (2018). Sex and credit: Is there a gender bias in lending? *Journal of Banking and Finance* 87.
- Becker, G. S. (1957). The economics of discrimination (chicago: University of chicago).
- Bellucci, A., A. Borisov, and A. Zazzaro (2010). Does gender matter in bank–firm relationships? evidence from small business lending. *Journal of Banking & Finance* 34(12), 2968–2984.
- Berente, N., B. Gu, J. Recker, and R. Santhanam (2021). Managing artificial intelligence.
- Berkovec, J. A., G. B. Canner, S. A. Gabriel, and T. H. Hannan (1998). Discrimination, competition, and loan performance in fha mortgage lending. *Review of Economics and Statistics* 80(2), 241–250.
- Bertrand, M. and S. Mullainathan (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review* 94(4), 991–1013.
- Blanchflower, D. G., P. B. Levine, and D. J. Zimmerman (2003). Discrimination in the small-business credit market. *Review of Economics and Statistics* 85(4), 930–943.
- Blau, F. D. and L. M. Kahn (2017). The gender wage gap: Extent, trends, and explanations. *Journal of economic literature* 55(3), 789–865.
- Bohren, J. A., K. Haggag, A. Imas, and D. G. Pope (2023). Inaccurate statistical discrimination: An identification problem. *Review of Economics and Statistics*, 1–45.
- Bohren, J. A., A. Imas, and M. Rosenberg (2019). The dynamics of discrimination: Theory and evidence. *American economic review* 109(10), 3395–3436.
- Bowen III, D. E., S. M. Price, L. C. Stein, and K. Yang (2024). Measuring and mitigating racial bias in large language model mortgage underwriting. *Available at SSRN* 4812158.
- Brock, J. M. and R. De Haas (2023). Discriminatory lending: Evidence from bankers in the lab. *American Economic Journal: Applied Economics* 15(2), 31–68.
- Bryan, G., D. Karlan, and A. Osman (2024). Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment. *American Economic Review* 114(9), 2825–2860.
- Brynjolfsson, E., D. Li, and L. Raymond (2025). Generative ai at work. *The Quarterly Journal of Economics* 140(2), 889–942.

- Buolamwini, J. and T. Gebru (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR.
- Butler, A. W., E. J. Mayer, and J. P. Weston (2023). Racial disparities in the auto loan market. *The Review of Financial Studies* 36(1), 1–41.
- Carter, S., E. Shaw, W. Lam, and F. Wilson (2007). Gender, entrepreneurship, and bank lending: The criteria and processes used by bank loan officers in assessing applications. *Entrepreneurship theory and practice* 31(3), 427–444.
- Cavalluzzo, K. S. and L. C. Cavalluzzo (1998). Market structure and discrimination: The case of small businesses. *Journal of Money, Credit and Banking*, 771–792.
- Cavalluzzo, K. S., L. C. Cavalluzzo, and J. D. Wolken (2002). Competition, small business financing, and discrimination: Evidence from a new survey. *The Journal of Business* 75(4), 641–679.
- Charles, K. K. and E. Hurst (2002). The transition to home ownership and the black-white wealth gap. *Review of Economics and Statistics* 84(2), 281–297.
- Charles, K. K., E. Hurst, and M. Stephens Jr (2008). Rates for vehicle loans: race and loan source. *American Economic Review* 98(2), 315–320.
- Cheraghali, H. and P. Molnár (2024). Sme default prediction: A systematic methodology-focused review. *Journal of Small Business Management* 62(6), 2847–2905.
- Chisadza, C., M. Clance, N. Nicholls, E. Yitbarek, and T. Zawaira (2025). Attitudes to gender inequality in south africa: Evidence from implicit and explicit attitudes.
- Cohen-Cole, E. (2011). Credit card redlining. *Review of Economics and Statistics* 93(2), 700–713.
- Cole, S., M. Kanz, and L. Klapper (2015). Incentivizing calculated risk-taking: Evidence from an experiment with commercial bank loan officers. *The Journal of Finance* 70(2), 537–575.
- Cook, T. R. and S. Kazinnik (2025). Social group bias in ai finance. *arXiv preprint arXiv:2506.17490*.
- Costello, A. M., A. K. Down, and M. N. Mehta (2020). Machine+ man: A field experiment on the role of discretion in augmenting ai-based lending models. *Journal of Accounting and Economics* 70(2-3), 101360.
- Crépon, B., M. El Komi, and A. Osman (2024). Is it who you are or what you get? comparing the impacts of loans and grants for microenterprise development. *American Economic Journal: Applied Economics* 16(1), 286–313.

- Dargnies, M.-P., R. Hakimov, and D. Kübler (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.
- Dastin, J. (2022). Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of data and analytics*, pp. 296–299. Auerbach Publications.
- De Mel, S., D. McKenzie, and C. Woodruff (2009). Are women more credit constrained? experimental evidence on gender and microenterprise returns. *American Economic Journal: Applied Economics* 1(3), 1–32.
- Deku, S. Y., A. Kara, and P. Molyneux (2016). Access to consumer credit in the uk. *The European Journal of Finance* 22(10), 941–964.
- Dell’Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Kraymer, F. Candelon, and K. R. Lakhani (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (24-013).
- Demirguc-Kunt, A., L. Klapper, D. Singer, S. Ansar, and J. Hess (2018). *The Global Findex Database 2017: Measuring financial inclusion and the fintech revolution*. World Bank Publications.
- Dobbie, W., A. Liberman, D. Paravisini, and V. Pathania (2021). Measuring bias in consumer lending. *The Review of Economic Studies* 88(6), 2799–2832.
- Egan, M., G. Matvos, and A. Seru (2022). When harry fired sally: The double standard in punishing misconduct. *Journal of Political Economy* 130(5), 1184–1248.
- Elam, A. B., K. D. Hughes, M. Guerrero, S. Hill, C. Nawangpalupi, M. Fuentes, and J. P. D. González (2021). Women’s entrepreneurship 2020/21: Thriving through crisis. *London: Global Entrepreneurship Research Association, London Business School*. Retrieved January 20, 2022.
- Etgar, S., G. Oestreicher-Singer, and I. Yahav (2024). Implicit bias in llms: bias in financial advice based on implied gender. *Available at SSRN*.
- Eyraud, L., I. Bunda, J. Jack, M. T. Jardak, R. Ouedraogo, Z. Wang, and T. Wezel (2021). *Resolving nonperforming loans in Sub-Saharan Africa in the Aftermath of the COVID-19 Crisis*. International Monetary Fund.
- FDIC (2024). Consumer compliance examination manual. Accessed June 4, 2025.

- Fisman, R., D. Paravisini, and V. Vig (2017). Cultural proximity and loan outcomes. *American Economic Review* 107(2), 457–492.
- Flage, A. (2018). Ethnic and gender discrimination in the rental housing market: Evidence from a meta-analysis of correspondence tests, 2006–2017. *Journal of Housing Economics* 41, 251–273.
- Foroughifar, M., R. Ghanavi, A. Goldfarb, and R. Webb (2025, September). Communicating uncertainty can increase ai adoption. Available at SSRN. SSRN working paper.
- Frame, W. S., R. Huang, E. X. Jiang, Y. Lee, W. S. Liu, E. J. Mayer, and A. Sunderam (2025). The impact of minority representation at mortgage lenders. *The Journal of Finance* 80(2), 1209–1260.
- Fu, R., Y. Huang, and P. V. Singh (2021). Crowds, lending, machine, and bias. *Information Systems Research* 32(1), 72–92.
- Fumagalli, E., S. Rezaei, and A. Salomons (2022). Ok computer: Worker perceptions of algorithmic recruitment. *Research Policy* 51(2), 104420.
- Fuster, A., P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther (2022). Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance* 77(1), 5–47.
- Glick, P. and S. T. Fiske (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology* 70(3), 491–512.
- Glover, D., A. Pallais, and W. Pariente (2017). Discrimination as a self-fulfilling prophecy: Evidence from french grocery stores. *The Quarterly Journal of Economics* 132(3), 1219–1260.
- Goh, E., R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, et al. (2024). Influence of a large language model on diagnostic reasoning: A randomized clinical vignette study. *medRxiv*.
- Goldin, C. (2014). A grand gender convergence: Its last chapter. *American economic review* 104(4), 1091–1119.
- Grand-Clément, J. and J. Pauphilet (2024). The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science*.
- Greenwald, A. G., D. E. McGhee, and J. L. Schwartz (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology* 74(6), 1464.

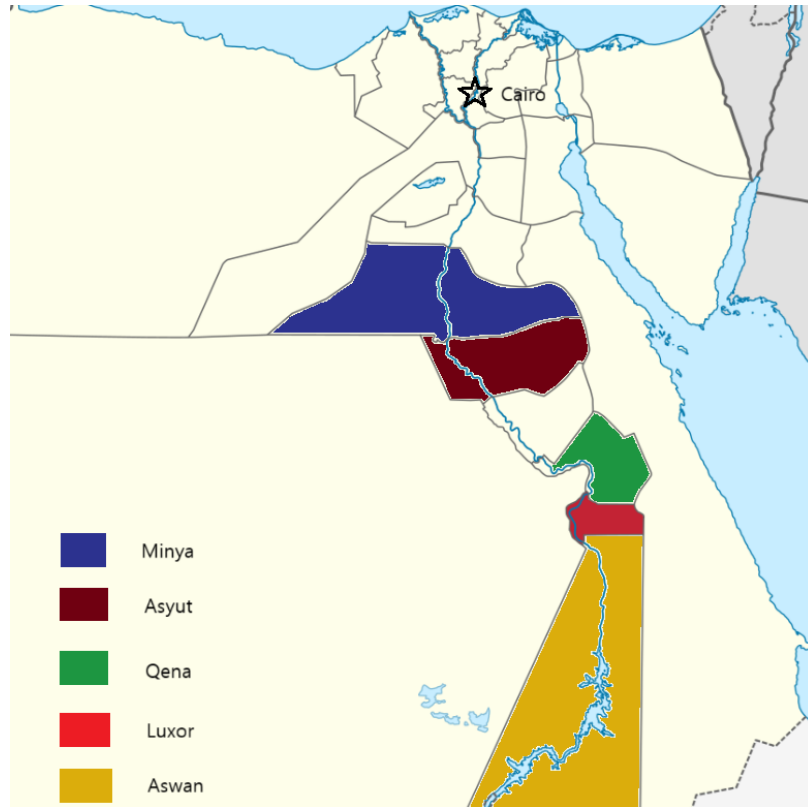
- Greenwald, A. G., B. A. Nosek, and M. R. Banaji (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology* 85(2), 197.
- Haerpfer, C., R. Inglehart, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin, B. Puranen, et al. (2022). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat* 12(10), 8.
- Han, S. (2004). Discrimination in lending: Theory and evidence. *The Journal of Real Estate Finance and Economics* 29(1), 5–46.
- Hanson, A., Z. Hawley, H. Martin, and B. Liu (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics* 92, 48–65.
- Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic literature* 42(4), 1009–1055.
- Haselmann, R., D. Schoenherr, and V. Vig (2018). Rent seeking in elite networks. *Journal of Political Economy* 126(4), 1638–1690.
- Hays, R., T. Hayashi, and A. Stewart (1989a). Socially desirable response set five-item survey (sdrs-5). *Educ Psychol Measur* 49, 629–39.
- Hays, R. D., T. Hayashi, and A. L. Stewart (1989b). A five-item measure of socially desirable response set. *Educational and psychological measurement* 49(3), 629–636.
- Hebert, C. (2023). Gender stereotypes and entrepreneur financing. In *10th Miami behavioral finance conference*.
- Hebert, C., H. Tookes, and E. Yimfor (2024). Financing the next vc-backed startup: The role of gender. *Available at SSRN*.
- Heywood, J. S. and P. L. O’Halloran (2005). Racial earnings differentials and performance pay. *Journal of Human Resources* 40(2), 435–452.
- Heywood, J. S. and D. Parent (2012). Performance pay and the white-black wage gap. *Journal of Labor Economics* 30(2), 249–290.
- Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The allocation of talent and us economic growth. *Econometrica* 87(5), 1439–1474.
- Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *The Quarterly journal of economics* 124(4), 1403–1448.

- Hu, X., Y. Huang, B. Li, and T. Lu (2025). Human–algorithmic bias: Source, evolution, and impact. *Management Science*.
- Hurlin, C., C. Pérignon, and S. Saurin (2024). The fairness of credit scoring models. *Management Science*.
- Kessler, J. B., C. Low, and C. D. Sullivan (2019). Incentivized resume rating: Eliciting employer preferences without deception. *American Economic Review* 109(11), 3713–3744.
- Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The quarterly journal of economics* 133(1), 237–293.
- Lane, K. A., M. R. Banaji, B. A. Nosek, and A. G. Greenwald (2007). Understanding and using the implicit association test: Iv: What we know (so far) about the method.
- Marinescu, I., N. Klein, A. Chamberlain, and M. Smart (2018). Incentives can reduce bias in online reviews. Technical report, National Bureau of Economic Research.
- Mascia, D. V. and S. P. Rossi (2017). Is there a gender effect on the cost of bank financing? *Journal of Financial Stability* 31, 136–153.
- Mejia, J. and C. Parker (2021). When transparency fails: Bias and financial incentives in ridesharing platforms. *Management Science* 67(1), 166–184.
- Mengel, F., J. Sauermann, and U. Zölitz (2019). Gender bias in teaching evaluations. *Journal of the European economic association* 17(2), 535–566.
- Mocanu, T. (2022). Designing gender equity: Evidence from hiring practices and committees. Technical report, Working paper.
- Moll, B. (2014). Productivity losses from financial frictions: Can self-financing undo capital misallocation? *American Economic Review* 104(10), 3186–3221.
- Montoya, A. M., E. Parrado, A. Solis, and R. Undurraga (2024). Bad taste: Gender discrimination in consumer lending.
- Muravyev, A., O. Talavera, and D. Schäfer (2009). Entrepreneurs’ gender and financial constraints: Evidence from international data. *Journal of comparative economics* 37(2), 270–286.
- Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature* 56(3), 799–866.

- Noy, S. and W. Zhang (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381(6654), 187–192.
- Osman, A., J. Speer, and A. Weaver (2023). Discrimination against women in hiring.
- Otis, N., R. P. Clarke, S. Delecourt, D. Holtz, and R. Koning (2023). The uneven impact of generative ai on entrepreneurial performance. *Available at SSRN* 4671369.
- Phelps, E. S. (1972). The statistical theory of racism and sexism. *The american economic review* 62(4), 659–661.
- Pope, D. G. and J. R. Sydnor (2011). What’s in a picture?: Evidence of discrimination from prosper. com. *Journal of Human resources* 46(1), 53–92.
- Rigol, N. and B. N. Roth (2021). Loan officers impede graduation from microfinance: Strategic disclosure in a large microfinance institution. Technical report, National Bureau of Economic Research.
- Ross, S. L., M. A. Turner, E. Godfrey, and R. R. Smith (2008). Mortgage lending in chicago and los angeles: A paired testing study of the pre-application process. *Journal of Urban Economics* 63(3), 902–919.
- Sarsons, H. (2017). Interpreting signals in the labor market: evidence from medical referrals. *Job market paper*, 141–145.
- Stefani, M. L. and V. P. Vacca (2013). Credit access for female firms: Evidence from a survey on european smes. *Bank of Italy Occasional Paper* (176).
- Stevenson, M. T. and J. L. Doleac (2024). Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy* 16(4), 382–414.
- Ubfal, D. (2024). What works in supporting women-led businesses?
- Yu, F., A. Moehring, O. Banerjee, T. Salz, N. Agarwal, and P. Rajpurkar (2024). Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nature Medicine* 30(3), 837–849.

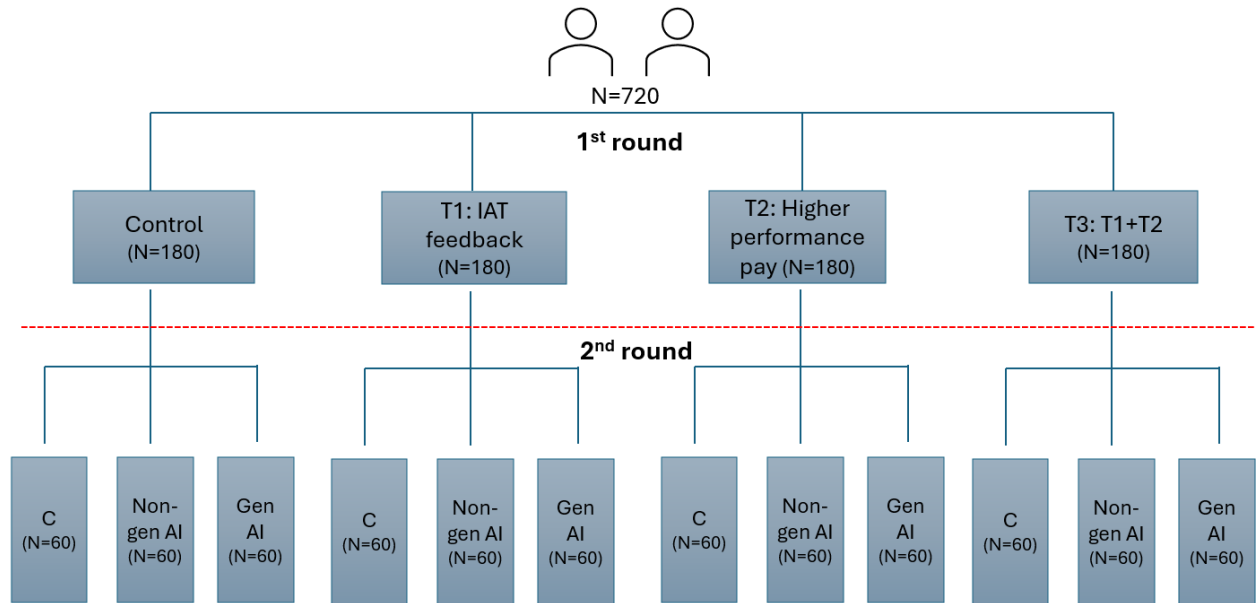
Figures

Figure 1. Geographical Distribution of Loan Officers by governorates



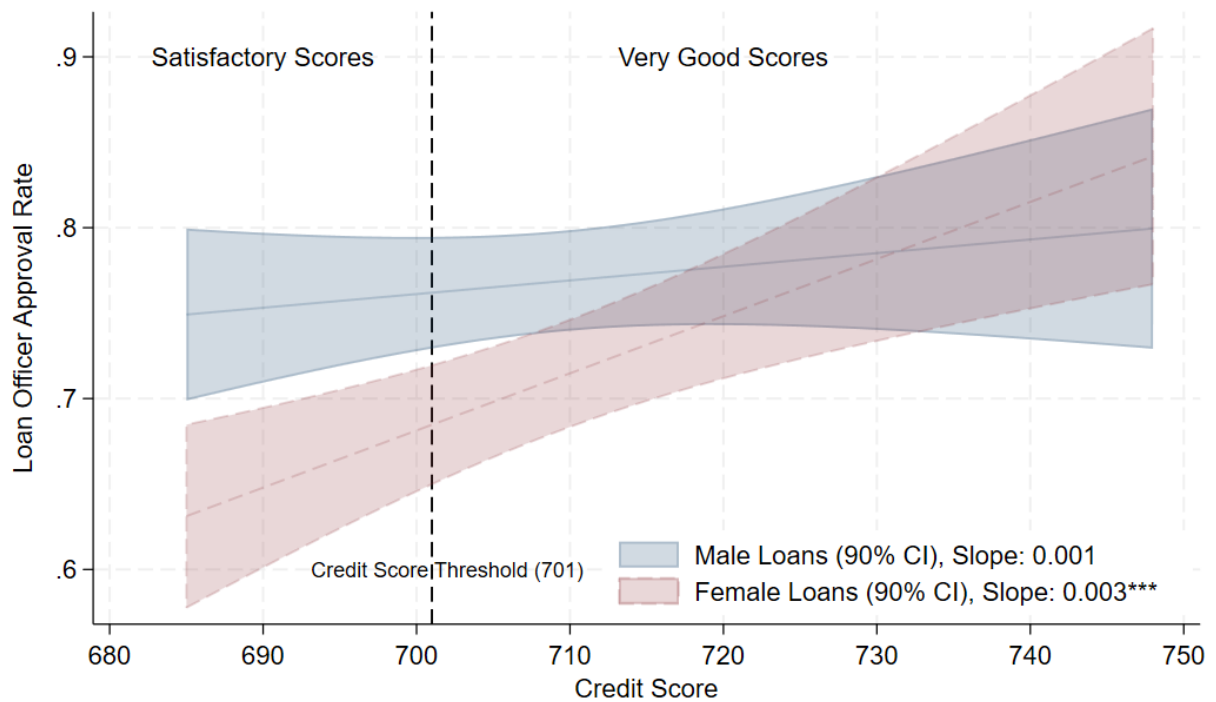
Notes: This figure shows geographical distribution of the five governorates where the experiment takes place—Aswan, Qena, Luxor, Asyut, and Minya. Egypt's capital, Cairo, is marked with a star. Source: Wikipedia: Luxor Governorate; map recolored by author.

Figure 2. Randomization Design



Notes: The randomization is conducted at the individual level. In the first round, T₃ refers to the combined treatment group, where officers receive both IAT feedback (T₁) and higher performance pay (T₂) together. In the second round, C stands for control group, Non-gen AI stands for non-generative AI group, and Gen-AI stands for generative AI group.

Figure 3. The Linear Relationship between Credit Score and Approval Rate by Gender of Applicants



Notes: This figure uses 31 out of 50 loans with available credit scores ranging from 685 to 750. Of the remaining 19 loans, 16 are missing credit scores, and 3 have low scores of 536, 537, and 622, respectively. These three loans are excluded from the figure because their credit scores are outliers compared to the typical range of 685 to 750, which could distort the overall trend. A credit score between 626 and 700 is considered satisfactory, while scores between 701 and 750 are classified as very good. The shaded area around the fitted line represents the 90% confidence interval. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Tables

Table 1: Baseline Balance Check

	(1)	(2)	(3)	(4)	(5)	(6)
	All	Control	IAT Feedback (Treatment1)	Higher Performance Pay (Treatment2)	Combined Intervention (Treatment3)	Joint F-test p-value
Female (=1)	0.50 {0.50}	0.50 {0.50}	0.50 {0.50}	0.50 {0.50}	0.50 {0.50}	1.00
Age	34.62 {8.22}	34.25 {7.83}	35.02 {8.57}	35.28 {8.79}	33.94 {7.62}	0.34
Married (=1)	0.66 {0.47}	0.70 {0.46}	0.66 {0.47}	0.66 {0.47}	0.63 {0.48}	0.52
Year of Education	16.20 {1.83}	16.34 {1.75}	16.14 {1.88}	16.01 {1.90}	16.30 {1.77}	0.31
Total Monthly Income (EGP)	6048 {3843}	6200 {3827}	5962 {3584}	6037 {3936}	5995 {4038}	0.93
Senior or above Rank LO Binary	0.29 {0.46}	0.29 {0.46}	0.34 {0.47}	0.28 {0.45}	0.26 {0.44}	0.41
Number of Experience (Year)	6.28 {5.72}	5.82 {5.45}	6.44 {5.90}	6.69 {6.20}	6.15 {5.28}	0.47
Household Size	3.70 {1.68}	3.62 {1.68}	3.70 {1.67}	3.68 {1.67}	3.80 {1.72}	0.80
Monthly Approval Rate	67.33 {16.90}	67.09 {16.05}	67.70 {17.23}	66.78 {16.79}	67.74 {17.62}	0.94
Loan Screen Time (hour)	10.51 {15.21}	11.05 {15.34}	12.03 {16.68}	9.56 {14.30}	9.41 {14.35}	0.31
LO Skill Binary (1=Soft, 0=Hard)	0.69 {0.46}	0.70 {0.46}	0.70 {0.46}	0.73 {0.45}	0.65 {0.48}	0.46
Weekly Working Hours	40.93 {11.12}	42.57 {11.87}	41.76 {11.10}	39.81 {10.97}	39.60 {10.28}	0.02
Accuracy rate in Cognitive Test	0.30 {0.23}	0.29 {0.24}	0.29 {0.23}	0.30 {0.23}	0.31 {0.23}	0.59
Digit Span Recall	3.98 {1.24}	4.00 {1.27}	3.93 {1.41}	4.07 {1.18}	3.92 {1.09}	0.59
Psychometric Test	3.65 {0.45}	3.63 {0.43}	3.64 {0.42}	3.66 {0.45}	3.65 {0.49}	0.89
Social Desirability Bias (Out of 100)	23.22 {22.76}	24.33 {23.47}	21.89 {22.04}	24.78 {22.90}	21.89 {22.64}	0.45
IAT Experience Before (=1)	0.14 {0.35}	0.16 {0.36}	0.16 {0.37}	0.11 {0.31}	0.14 {0.35}	0.37
IAT Score	0.37 {0.53}	0.40 {0.49}	0.32 {0.54}	0.41 {0.55}	0.33 {0.54}	0.26
Explicit Gender Bias (Business)	2.82 {1.26}	2.80 {1.24}	2.72 {1.28}	2.93 {1.26}	2.82 {1.27}	0.47
Explicit Gender Bias (All)	2.98 {0.73}	2.98 {0.70}	2.90 {0.71}	3.05 {0.72}	2.98 {0.78}	0.30
Risk Preference (Out of 10)	6.44 {2.95}	6.49 {2.97}	6.11 {3.08}	6.73 {2.88}	6.42 {2.86}	0.25
Investment Choice on Fair 50/50 Chance (Out of 100)	52.30 {24.21}	51.53 {24.75}	50.92 {22.61}	54.15 {25.61}	52.61 {23.86}	0.61
Gender Education Experience Binary	0.44 {0.50}	0.41 {0.49}	0.46 {0.50}	0.46 {0.50}	0.43 {0.50}	0.78
Existence of Second Job Binary	0.20 {0.40}	0.19 {0.39}	0.17 {0.37}	0.22 {0.41}	0.24 {0.43}	0.34
Joint F-Test p-value						0.404
Number of Observations	720	180	180	180	180	720

Notes: This table shows balance check of baseline variables by first-round treatment arms. Means are presented with standard deviations shown in brackets below. The monetary variables are winsorized by top and bottom 1%. The details of the baseline characteristic variables are presented in Appendix D.

Table 2: Logit Regression Results (Ex-post Defaulted vs. Repaid Loans)

	Ex-post Defaulted Loans
log (Loan Size)	3.31* (1.88)
Loan Tenure (Months)	0.31*** (0.07)
log (Monthly Income)	0.71 (3.85)
Late Repayment History Binary	1.71* (0.99)
Number of Previous Loans	0.14 (0.26)
Number of Employees	-2.98** (1.49)
Registered Business Binary	-2.52 (2.66)
Age	-0.05 (0.07)
Number of Employee Guarantors	-3.09 (2.26)
Joint F-test Pvalue	0.001
Number of Loans	50

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the coefficients from a logistic regression that compares ex-post defaulted loans to ex-post repaid loans. Standard errors are reported in parentheses. Among the 50 loans analyzed, 30 were repaid and 20 defaulted. The regression includes 9 continuous loan characteristics, excluding categorical variables such as industry and region.

Table 3: Approval Rate Gap by Gender of Portfolio: Ex-post Outcome

	(1) Total Loans	(2) Ex-post Defaulted Loans (Generous to Men)	(3) Ex-post Repaid Loans (Harsh to Women)
Male Names	0.044** (0.019)	0.097*** (0.032)	0.009 (0.022)
Female Name Approval Mean	0.708	0.661	0.739
Number of Portfolios	1800	720	1080
Number of Loan Officers	180	180	180
Portfolio FE	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the gender gap in loan approvals for all loans, ex-post repaid loans, and ex-post defaulted loans. Ex-post defaulted loans refer to those with repayment delays of 90 days or more. The outcome is a binary variable equal to 1 if the loan officer approves a loan, and 0 if they reject it. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 4-1. Approval Rate Gap by Gender of Portfolio: Late Repayment History

	(1) Loans with Late Repayment History (Generous to Men)	(2) Loans without Late Repayment History (Harsh to Women)
Male Names	0.111** (0.047)	0.026 (0.021)
Female Name Approval Mean	0.515	0.762
T-test p-value (col1 vs. col2)		0.094
Number of Portfolios	396	1404
Number of Loan Officers	180	180
Portfolio FE	✓	✓
Loan Officer Gender FE	✓	✓
Cohort FE	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The late repayment history indicates whether the loan had at least one day of overdue repayment in the most recent loan. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 4-2. Approval Rate Gap by Gender of Portfolio: Credit Score

	(1) Satisfactory or Lower Credit Score (700 or Below) :Generous to Men	(2) Very Good Credit Score (701 or Above) : Harsh to Women	(3) Credit Score Not Available
Male Names	0.097*** (0.037)	0.006 (0.031)	0.035 (0.034)
Female Name Approval Mean	0.576	0.778	0.760
T-test p-value (col1 vs. col2)		0.058	.
Monthly repayment/Income Mean	0.404	0.292	0.287
Number of Portfolios	576	648	576
Number of Loan Officers	180	180	180
Portfolio FE	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Column 1 presents the subgroup analysis for loans with an available credit score 700 or below. Column 2 includes loans with a credit score of 701 or above, and Column 3 covers loans without an available credit score. The variable monthly repayment/income refers to the ratio of monthly repayment amount to monthly income, with higher values indicating riskier loans. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 5. Debiasing Treatment Effects in Loan Approval

	(1)	(2)	(3)
	All Officers	Officers Biased Against Women (IAT Score ≥ 0.35)	Others (IAT Score < 0.35)
IAT Feedback (T1) x Male Names	-0.039 (0.026)	-0.026 (0.039)	-0.040 (0.037)
Higher Performance Pay (T2) x Male Names	-0.066** (0.027)	-0.080** (0.035)	-0.019 (0.043)
Combined (T3) x Male Names	-0.048* (0.027)	-0.064* (0.037)	-0.017 (0.040)
Male Names	0.044** (0.019)	0.047* (0.026)	0.028 (0.030)
T-test p-value (T2 vs. T1)	0.478	0.306	0.707
T-test p-value (T2 vs. T3)	0.640	0.764	0.970
Number of Portfolios	7200	3790	3300
Number of Loan Officers	720	379	330
Portfolio FE	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the change in the gender gap in loan approval rates for each treatment arm relative to the control group. The combined group refers to the treatment group 3, where officers receive both the IAT feedback and the higher performance pay. Column (1) reports results for all loan officers. Column (2) shows results for officers with a Gender-Career IAT score of 0.35 or higher, indicating a medium or stronger association between men and careers and women and family. Column (3) reports results for officers with an IAT score below 0.35, indicating a weaker, no, or opposite association. Due to a technical issue, IAT scores were not recorded for 11 officers in the control and higher-performance-pay groups; these observations are excluded from Columns (2) and (3). Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table 6. Debiasing Treatment Effects in Loan Screening Performance and Time

	(1)	(2)	(3)
	Accuracy (Male-name Loans)	Accuracy (Female-name Loans)	Average Screening Time per Loan (minutes)
IAT Feedback (T1)	0.021 (0.019)	-0.004 (0.018)	-0.033 (0.045)
Higher Performance Pay (T2)	0.036* (0.020)	-0.021 (0.019)	0.124** (0.052)
Combined (T3)	0.042** (0.019)	-0.001 (0.018)	0.006 (0.044)
Control Group Mean	0.546	0.579	1.098
T-test p-value (T2 vs. T1)	0.601	0.523	0.023
T-test p-value (T2 vs. T3)	0.809	0.445	0.083
Number of Portfolios	3600	3600	.
Number of Loan Officers	720	720	704
Portfolio FE	✓	✓	
Loan Officer Gender FE	✓	✓	✓
Cohort FE	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table reports changes in first-round loan screening performance (accuracy) and screening time for each treatment arm relative to the control group. The combined group refers to the treatment group 3, where officers receive both the IAT feedback and the higher performance pay. Screening performance is a binary variable equal to 1 if an ex-post repaid loan is approved or an ex-post defaulted loan is rejected. It is equal to 0 if an ex-post defaulted loan is approved or an ex-post repaid loan is rejected. Screening time is measured as the total duration spent reviewing all 10 loan applications; therefore, it is not possible to calculate screening time separately for male and female applicants. For the first 16 loan officers in the experiment, screening time was not recorded due to a technical error, and their observations are excluded from the Column (3) analysis. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 7. AI-assisted Decision and Loan Screening Performance

	(1) Combined AI's Effects	(2) Generative AI vs. Non-Generative AI
All AI Types (Generative + Non-Generative)	0.078*** (0.012)	
Generative AI		0.066*** (0.013)
Non-Generative AI		0.089*** (0.013)
2nd Control Accuracy Mean	0.553	0.553
AI Chatbot Accuracy	0.800	0.800
One Side P-value (Gen AI > Non-Gen AI)		0.115
Number of Portfolios	7200	7200
Number of Loan Officers	720	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the improvement in second-round loan screening performance (accuracy) resulting from the use of AI. Screening performance is a binary variable equal to 1 if an ex-post repaid loan is approved or an ex-post defaulted loan is rejected. It is equal to 0 if an ex-post defaulted loan is approved or an ex-post repaid loan is rejected. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table 8. Selective Follow Rates of AI Recommendations: Generative vs. Non-Generative AI

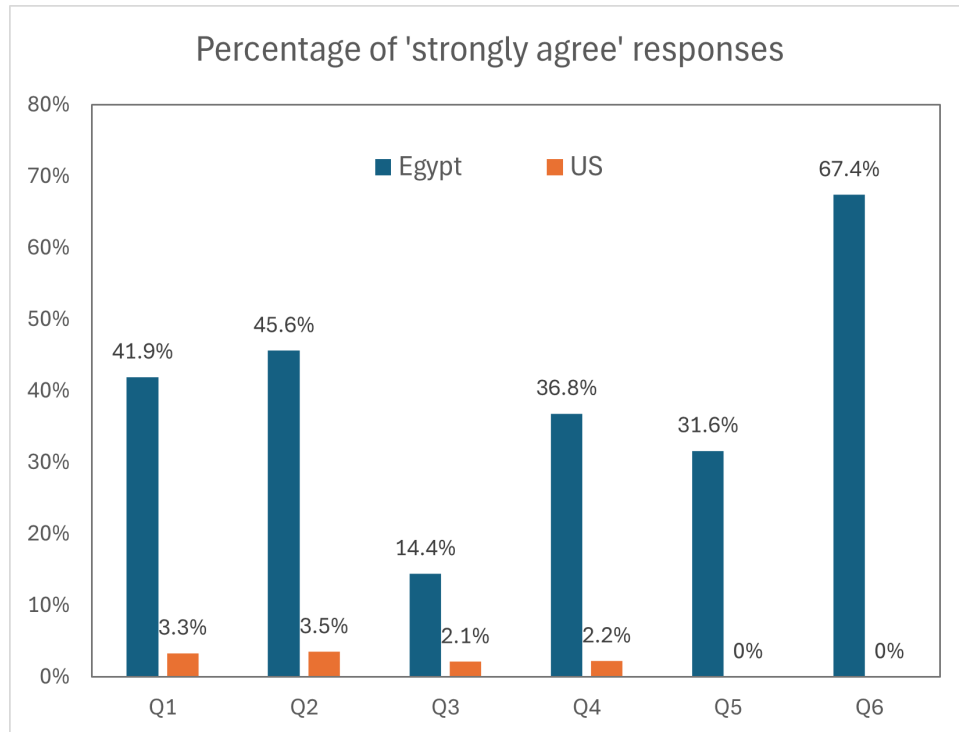
	Ex-post Defaulted Loans		Ex-post Repaid Loans
	(1) AI Decisions : Reject	(2) AI Decisions : Approve	(3) AI Decisions : Approve
Gen AI Binary x Male Names	0.025 (0.056)	0.113*** (0.042)	-0.021 (0.025)
Gen AI Binary	-0.012 (0.041)	-0.063* (0.033)	-0.029 (0.021)
Male Names	0.013 (0.038)	-0.025 (0.030)	0.024 (0.016)
Number of Portfolios	960	960	2880
Number of Loan Officers	480	480	480

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows loan officers' follow rate for AI recommendations in the second round. The AI is programmed to make three types of recommendations: approve ex-post repaid loans, approve ex-post defaulted loans, and reject ex-post defaulted loans. The outcome is a binary follow-rate variable, equal to 1 if the officer follows the AI's recommendation and 0 otherwise. Standard errors, clustered at the loan officer level, are reported in parentheses.

7 Appendix Figures and Tables

Figure A1. Response to 6 gender bias questions in the 2017 World Value Survey



Notes: This figure compares the percentage of "Strongly Agree" responses between the United States and Egypt in the 2017 World Value Survey. The questions are: (1) When a mother works for pay, the children suffer. (2) On the whole, men make better political leaders than women do. (3) A university education is more important for a boy than for a girl. (4) On the whole, men make better business executives than women do. (5) If a woman earns more money than her husband, it's almost certain to cause problems. (6) When jobs are scarce, men should have more right to a job than women.

Figure A2. Illustrative Example of a Loan Application in the Experiment

Client Name	<i>Ahmed Mohamed</i>	Age	<i>30</i>	Number of Employee Guarantor	<i>2</i>
Number of Non- Employee Guarantor	<i>1</i>	Monthly income	<i>10,000 EGP</i>	Registered business Binary	<i>Yes</i>
Business start year	<i>2015</i>	Location	<i>Aswan city</i>	Number of employees	<i>3</i>
Industry	<i>Agriculture</i>	Loan size	<i>10,000 EGP</i>	Loan term	<i>24 months</i>
Interest rate	<i>20%</i>	Number of previous loans	<i>3</i>	Late Repayment Days late on most recent loan	<i>0</i>
Loan size on most recent loan	<i>10,000 EGP</i>	Interest rate (%) on most recent loan	<i>20%</i>	Credit score	<i>700</i>

Notes: This figure presents a pseudo-example of one portfolio used in the experiment. All information has been replaced with fictitious data and does not contain any personal details.

Figure A3. Example of IAT Feedback Provided to the IAT Feedback (T1) and the Combined Group (T3)

Thanks for your participation. Your Gender-Career IAT score is Here's how to understand your results:

If your score is positive, it indicates a stronger subconscious association between "male" and "career" or "female" and "family." Conversely, if your score is negative, it signifies a stronger subconscious association between "male" and "family" or "female" and "career." Here's how to interpret the magnitude of your score:

No association if the score is between -0.15 and 0.15 , slight association for values between $|0.15|$ and $|0.35|$, moderate association between $|0.35|$ and $|0.65|$, and strong association for scores higher than $|0.65|$.

We want to iterate that this test reveals implicit attitudes and not behaviors. Our attitudes may derive from the cultural and social context where we live. An enormous body of literature confirms that we all have biases—some explicit, many implicit. However, it is important to avoid our implicit biases or stereotypes related to a specific group from systematically influencing our behavior toward clients. Acknowledging and understanding our biases and those of our colleagues can help minimize the influence they have on our daily interaction with clients.

← Previous

Next →

Notes: This figure shows an example of the IAT feedback provided to the IAT Feedback group (T1) and the Combined group (T3). The format is adapted from the widely used, publicly available version at <https://implicit.harvard.edu/implicit/>. In contrast, the Control group (C) and the Higher Performance Pay group (T2) received only a simple message after completing the IAT test: Thanks for your participation.

Figure A4. User Interface Viewed by Non-Generative AI Users

Number of guarantors who are employees: 0
Number of guarantors who are not employees: 1
Industry in which the client's business operates: Poultry
Whether the company is registered with the government (1=Yes, 0=No): 0
Year of company establishment: 2014
Area of business in Egypt: Edfu, Aswan
Number of employees in the company: 1
Amount of loan granted: 15,000
Loan term in months: 18 months
Current interest rate on the loan: 24.0
Number of previous loans obtained by the client: 3
Size of previous loan: 40,000.0
Number of days late in previous loan: 1.0
Interest rate on previous loan: 16.0
Client credit score: 696.0
Loan Decision: Approval

The decision to approve the loan for Nouha Ahmed could be based on the following reasons:

1. **Good Credit History** : The customer's credit score is 696, which is generally considered good. This indicates that the customer has a solid credit history and the ability to repay loans on time.
2. **Positive repayment history on previous loans** : Although the customer had previously borrowed a total of EGP 40,000, the number of days past due was only one. This indicates that the customer was generally committed to repaying loans on time, reducing the likelihood of being 90 days or more late in the future.

return the next

Notes: This figure displays a pseudo-example of the user interface that non-generative AI users interact with the AI chatbot. First, loan officers enter their randomized loan officer key on the screen, as shown in the figure above. If the loan officer's key corresponds to non-generative AI, the above screen appears on the screen. This figure shows the profile of the fictitious first portfolio in the bundle, along with the AI's decisions and the corresponding reasons. Questions are not allowed, and loan officers can navigate to the next or previous profile by clicking the buttons at the bottom. After reviewing the final portfolio, the program ends.

Figure A5. User Interface Viewed by Generative AI Users

chat box | interest rate on the loan: 24.0

Number of previous loans obtained by the client: 2
 Size of previous loan: 50,000.0
 Number of days late in previous loan: 0.0
 Interest rate on previous loan: 24.0
 Client credit score: 536.0
 Loan Decision: Approval

The decision to approve a loan for Ashraf Ibrahim could be based on the following reasons:

- 1. Good credit history on previous loans :** The customer has a history of obtaining loans and has not recorded any delinquent payments (number of days delinquent on previous loans: 0). This indicates their commitment to repaying loans on time, reducing the likelihood of future delinquencies.
- 2. A stable and reasonable monthly income :** The average monthly income of the client is EGP 20,000, which is considered a reasonable amount for the requested loan amount (EGP 75,000) and the short repayment term (6 months). This income can support the client's ability to repay the monthly loan installments without significant difficulty.

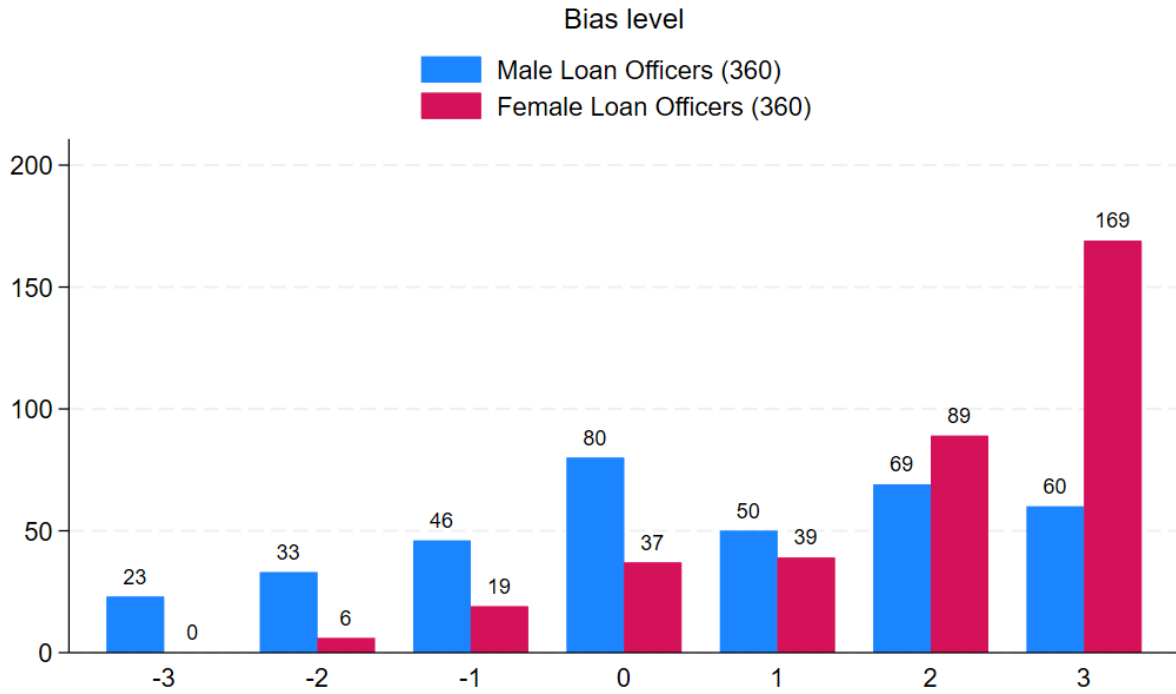
In addition, having a guarantor who is employed and a government-registered company strengthens the customer's position and increases confidence in their ability to repay.

If you have additional questions, you can ask them. Otherwise, click Next.

return the next

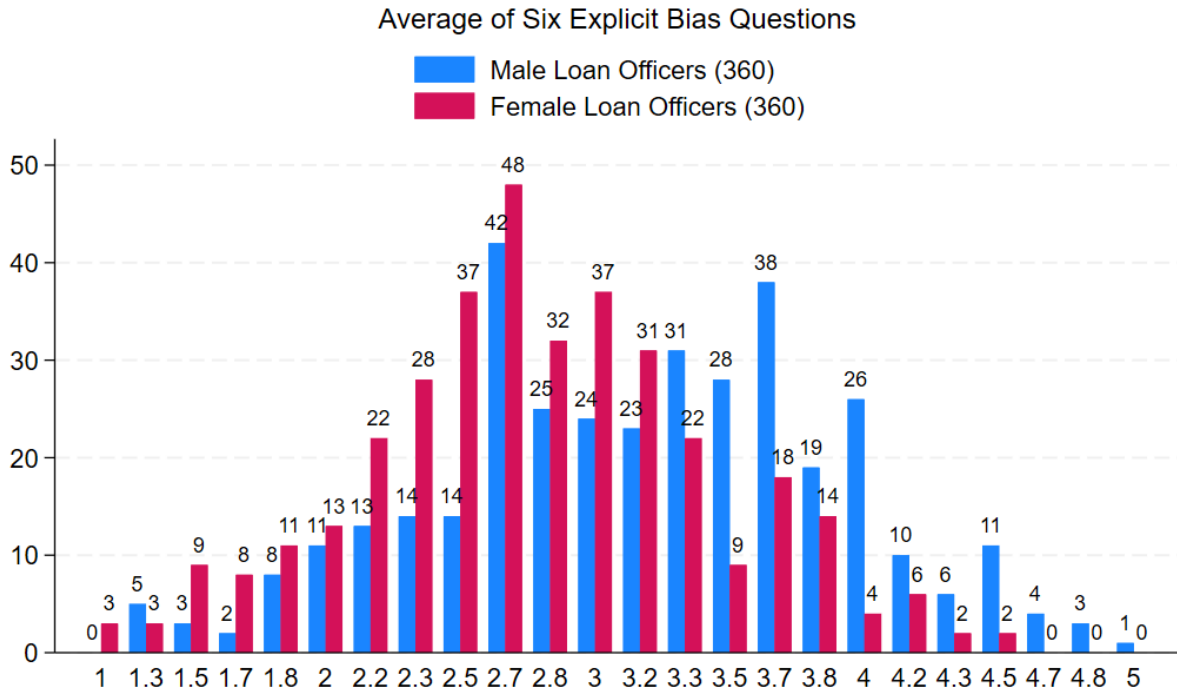
Notes: This figure displays a pseudo-example of the user interface that generative AI users interact with the AI chatbot. Similar to non-generative AI users, loan officers enter their loan officer key, and if it corresponds to the generative AI group, the generative AI interface appears. This figure shows the profile of the fictitious first portfolio in the bundle, along with the AI's decisions and the corresponding reasons. Unlike the non-generative AI interface, questions are allowed on the right side. If loan officers ask questions to the AI, the answers will appear below the questions on the right side. Loan officers can navigate to the next or previous profile by clicking the buttons at the bottom. After reviewing the final portfolio, the program ends.

Figure A6. The range of Gender-Career IAT score of loan officers



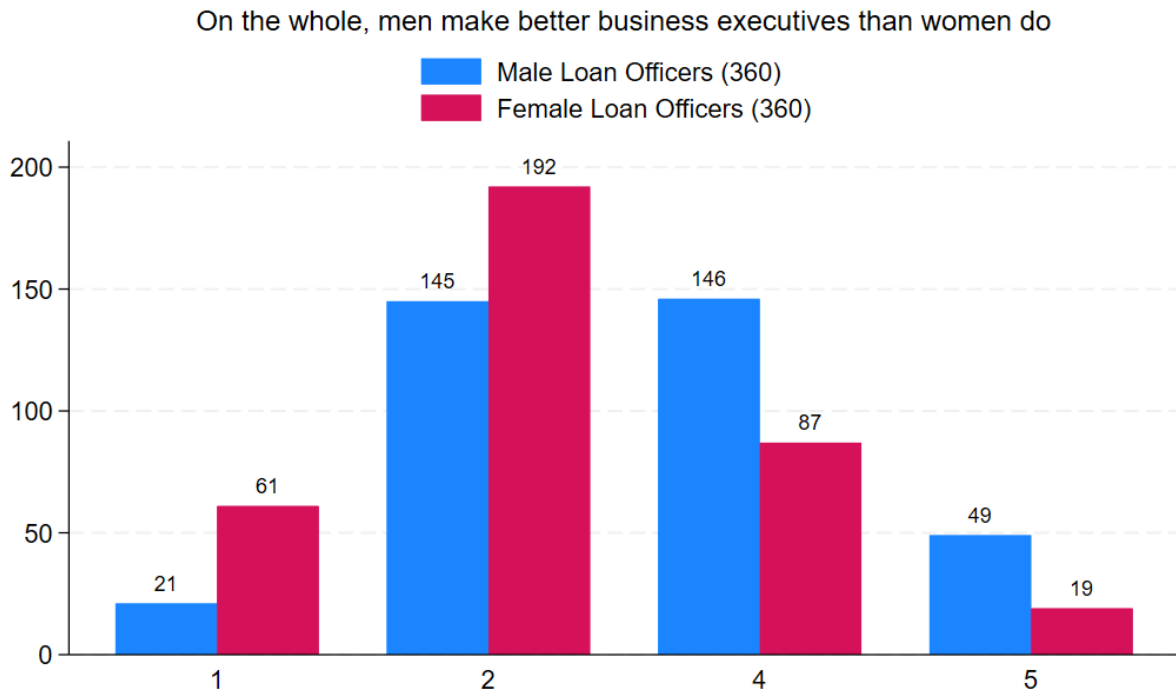
Notes: This figure shows the Gender-Career IAT score of 360 male loan officers and 360 female loan officers. Gender-Career IAT score labeling: -3: Strong bias against men ($IAT \leq -0.65$), -2: Moderate bias against men ($-0.65 < IAT \leq -0.35$), -1: Weak bias against men ($-0.35 < IAT \leq -0.15$), 0: No gender bias ($-0.15 < IAT < 0.15$), 1: Weak bias against women ($0.15 \leq IAT < 0.35$), 2: Moderate bias against women ($0.35 \leq IAT < 0.65$), 3: Strong bias against women ($IAT > 0.65$).

Figure A7. The average responses to explicit gender bias questions



Notes: This figure compares the average responses on a 5-point Likert scale for six explicit gender bias questions derived from the 2017 World Values Survey. The Likert scale ranges from 1 (strongly disagree) to 5 (strongly agree), with 3 representing a neutral response. The questions are: (1) When a mother works for pay, the children suffer. (2) On the whole, men make better political leaders than women do. (3) A university education is more important for a boy than for a girl. (4) On the whole, men make better business executives than women do. (5) If a woman earns more money than her husband, it's almost certain to cause problems. (6) When jobs are scarce, men should have more right to a job than women.

Figure A8. The response for an explicit bias question about business performance



Notes: This figure compares responses to an explicit gender bias question regarding business leadership, derived from the 2017 World Values Survey. Responses were recorded on a 4-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree), with 3 omitted as a neutral option in accordance with the World Values Survey format. The question posed was: “On the whole, men make better business executives than women do.”

Table A1. Summary of Participating Microfinance Institutions (MFIs)

MFI Number	Governorate	Number of Loan Officers		
		Total	Female	Male
1	Luxor and Aswan	185	88	97
2	Luxor and Qena	152	79	73
3	Asyut	100	68	32
4	Minya	51	1	50
5	Aswan	24	21	3
6	Minya	21	5	16
7	Minya	19	14	5
8	Minya	17	4	13
9	Minya	16	5	11
10	Minya	13	8	5
11	Minya	11	8	3
12	Qena	11	5	6
13	Minya	10	3	7
14	Minya	9	5	4
15	Minya	9	5	4
16	Minya	9	5	4
17	Minya	9	0	9
18	Minya	8	3	5
19	Minya	6	6	0
20	Minya	6	0	6
21	Aswan	5	5	0
22	Minya	5	2	3
23	Minya	5	5	0
24	Minya	5	5	0
25	Aswan	4	4	0
26	Minya	4	3	1
27	Minya	3	2	1
28	Minya	2	1	1
29	Minya	1	0	1
Total		720	360	360

Notes: This table shows the number of loan officers participated in the experiment by MFIs with governorate and gender of loan officers. The specific names of MFIs are not disclosed due to the some MFIs did not want to disclose their names. The MFIs are ordered by the number of loan officers.

Table A2. Payoff structure by treatment arms

	Control Group IAT Feedback (Treatment 1)	Higher Performance Pay (Treatment 2) Combined Intervention (Treatment 3)
Basic payment	400 EGP	400 EGP
Approval of an ex-post repaid loan	10 EGP	20 EGP
Rejection of an ex-post defaulted loan	15 EGP	30 EGP
Approval of an ex-post defaulted loan	-15 EGP	-30 EGP
Rejection of an ex-post repaid loan	-10 EGP	-20 EGP

Notes: This table compares the incentive and penalty structure of loan screening by treatment arms. The minimum payment (when all answers are wrong) is 280 EGP, with 160 EGP for each group. The maximum payment (when all answers are correct) is 520 EGP, or 640 EGP. As of the end of 2024, 1 USD was exchanged for approximately 50 EGP.

Table A3. Baseline Balance Check (Round 2)

	(1)	(2)	(3)	(4)	(5)
	All	AI Control	Non-Generative AI	Generative AI	Joint F-test p-value
Female (=1)	0.50 {0.50}	0.50 {0.50}	0.50 {0.50}	0.50 {0.50}	1.00
Age	34.62 {8.22}	34.18 {7.97}	34.36 {8.29}	35.33 {8.37}	0.89
Married (=1)	0.66 {0.47}	0.65 {0.48}	0.62 {0.49}	0.72 {0.45}	0.74
Year of Education	16.20 {1.83}	16.35 {1.85}	16.11 {1.82}	16.13 {1.81}	0.35
Total Monthly Income (EGP)	5897 {3353}	5532 {3086}	5851 {3351}	6307 {3575}	0.32
Senior or above Rank LO Binary	0.29 {0.46}	0.27 {0.45}	0.30 {0.46}	0.30 {0.46}	0.70
Number of Experience (Year)	6.18 {5.44}	5.72 {5.41}	6.09 {5.40}	6.72 {5.49}	0.72
Household Size	3.63 {1.52}	3.54 {1.55}	3.73 {1.54}	3.63 {1.47}	0.41
Monthly Approval Rate	67.12 {15.90}	65.93 {17.14}	68.29 {15.60}	67.15 {14.87}	0.28
Loan Screen Time (hour)	10.28 {14.45}	11.50 {15.35}	10.89 {14.86}	8.45 {12.90}	0.63
LO Skill Binary (1=Soft, 0=Hard)	0.69 {0.46}	0.72 {0.45}	0.64 {0.48}	0.72 {0.45}	0.12
Weekly Working Hours	40.81 {9.65}	41.55 {9.64}	41.65 {9.89}	39.22 {9.26}	0.99
Accuracy rate in Cognitive Test	0.30 {0.23}	0.28 {0.25}	0.30 {0.23}	0.31 {0.22}	0.16
Digit Span Recall	3.98 {1.24}	4.01 {1.23}	3.93 {1.25}	4.00 {1.25}	0.48
Psychometric Test	3.65 {0.45}	3.67 {0.45}	3.63 {0.44}	3.64 {0.45}	0.48
Social Desirability Bias (Out of 100)	23.22 {22.76}	24.08 {23.11}	22.17 {21.85}	23.42 {23.36}	0.44
IAT Experience Before (=1)	0.14 {0.35}	0.14 {0.35}	0.12 {0.33}	0.16 {0.37}	0.44
IAT Score	0.37 {0.50}	0.34 {0.52}	0.37 {0.49}	0.39 {0.50}	0.33
Explicit Gender Bias (Business)	2.82 {1.26}	2.78 {1.31}	2.80 {1.24}	2.86 {1.25}	0.83
Explicit Gender Bias (All)	2.98 {0.73}	2.95 {0.75}	2.98 {0.71}	3.01 {0.72}	0.49
Risk Preference (Out of 10)	6.44 {2.95}	6.40 {3.03}	6.41 {2.89}	6.51 {2.93}	0.99
Investment Choice on Fair 50/50 Chance (Out of 100)	52.30 {24.21}	53.21 {23.39}	51.94 {23.05}	51.76 {26.15}	0.84
Gender Education Experience Binary	0.44 {0.50}	0.50 {0.50}	0.42 {0.49}	0.39 {0.49}	0.07
Existence of Second Job Binary	0.20 {0.40}	0.19 {0.39}	0.20 {0.40}	0.22 {0.41}	0.79
Prior AI Use Experience Binary	0.31 {0.46}	0.29 {0.46}	0.30 {0.46}	0.35 {0.48}	0.84
Joint F-Test p-value					0.787
Number of Observations	720	240	240	240	720

Notes: This table shows the balance check of baseline variables across the second-round treatment arms. Means are presented with standard deviations shown in brackets below. The monetary variables are winsorized by top and bottom 1%. The details of the baseline characteristic variables are presented in Appendix D.

Table A4. Approval Rate Gap by Gender of Portfolio: ML Outcomes

	Predicted Probability of Repayment by ML (0.5 threshold)		Predicted Probability of Repayment by ML (Bottom 40% vs. Top 60%)	
	(1) Optimal Decision: Reject (<0.50) Generous to Men	(2) Optimal Decision: Approve (>=0.50) Harsh to Women	(3) Optimal Decision: Reject (Bottom 40%) Generous to Men	(4) Optimal Decision: Approve (Top 60%) Harsh to Women
Male Names	0.068** (0.031)	0.031 (0.023)	0.061** (0.030)	0.033 (0.024)
Female Names Approval Mean	0.654	0.738	0.667	0.735
T-test p-value		0.342		0.232
Number of Portfolios	648	1152	720	1080
Number of Loan Officers	180	180	180	180
Portfolio FE	✓	✓	✓	✓
Loan Officer Gender FE	✓	✓	✓	✓
Cohort FE	✓	✓	✓	✓

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each loan has a predicted repayment probability generated by the GBM model. Columns 1 and 2 classify loans using a threshold of 0.5. Loans with predicted probabilities below 0.5 are considered high-risk and optimally rejected, while those with probabilities above 0.5 are considered low-risk and optimally approved. Columns 3 and 4 apply the same classification process but use a higher threshold of 0.54 at which loans are classified into the bottom 40% and top 60%. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A5. Debiasing Treatment Effects in Loan Screening Performance by ML Outcomes

	(1) Accuracy (Male-name Loans)	(2) Accuracy (Female-name Loans)
IAT Feedback (T1)	0.006 (0.018)	-0.002 (0.019)
Higher Performance Pay (T2)	-0.016 (0.018)	-0.017 (0.020)
Combined (T3)	0.007 (0.018)	-0.001 (0.019)
Control Group Accuracy Mean	0.592	0.597
Number of Portfolios	7200	7200
Number of Loan Officers	720	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each loan has a predicted repayment probability generated by the GBM model. The outcome is a binary screening performance (accuracy) variable that equals 1 if the officer approves loans with predicted repayment probabilities above the threshold or rejects loans below the threshold, and 0 if the officer approves loans below the threshold or rejects loans above the threshold. A threshold of 0.50 is used to measure accuracy. Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A6. Screening Performance Change of AI-Assisted Decisions by ML Outcomes

	Predicted Probability of Repayment by ML (0.50 threshold)
Generative AI	0.027** (0.013)
Non-Generative AI	0.052*** (0.013)
2nd Control Accuracy Mean	0.593
AI Chatbot Accuracy	0.760
One Side P-value (Gen AI > Non-Gen AI)	0.090
Number of Portfolios	7200
Number of Loan Officers	720

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: Each loan has a predicted repayment probability generated by the GBM model. The outcome is a binary screening performance (accuracy) variable that equals 1 if the officer approves loans with predicted repayment probabilities above the threshold or rejects loans below the threshold, and 0 if the officer approves loans below the threshold or rejects loans above the threshold. A threshold of 0.50 is used to measure accuracy. The Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A7. Selective Follow Rates of AI Approval Recommendations for High-risk Loans by ML Outcomes

	AI Recommendation: Approval		
	(1) Repayment Probability by ML (<0.4)	(2) Repayment Probability by ML (<0.5)	(3) Repayment Probability by ML (<0.6)
Gen AI Binary x Male Names	0.101** (0.050)	0.088** (0.042)	0.059 (0.036)
Gen AI Binary	-0.012 (0.042)	-0.033 (0.034)	-0.017 (0.030)
Male Names	-0.042 (0.035)	-0.029 (0.029)	-0.028 (0.025)
Number of Portfolios	672	960	1152
Number of Loan Officers	480	480	480

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: The outcome is a binary variable equal to 1 if the officer follows the AI's approval recommendation, and 0 otherwise. Each loan has a predicted repayment probability generated by the GBM model. High-risk loans are defined as those with predicted repayment probabilities below the approval threshold, with multiple thresholds used (0.4, 0.5, and 0.6). Standard errors, clustered at the loan officer level, are reported in parentheses.

Table A8. Portfolios that Officers Asked Questions to the Chatbot

	(1)	(2)	(3)	(4)
	All	Approval by AI (Ex-post Repaid)	Approval by AI (Ex-post Defaulted)	Rejection by AI (Ex-post Defaulted)
% Portfolios Questioned	0.23	0.20	0.19	0.37
% Portfolios with Multiple Interaction	0.07	0.07	0.08	0.07
Number of Portfolios	2400	1440	480	480
Number of Loan Officers	240	240	240	240

Notes: This table presents summary statistics of questions asked to a generative AI chatbot by loan officers. The chatbot is programmed to provide three types of recommendations: approval of ex-post repaid loans, approval of ex-post defaulted loans, and rejection of ex-post defaulted loans. The term multiple interaction refers to instances where loan officers ask a question, receive a response from the chatbot, and then follow up with an additional question.

Table A9. Types of Questions Asked to the Chatbot

	(1)	(2)	(3)	(4)
	All	Approval by AI (Ex-post Repaid)	Approval by AI (Ex-post Defaulted)	Rejection by AI (Ex-post Defaulted)
% Information-seeking / Neutral	0.40	0.43	0.52	0.28
% Weak disagreement	0.50	0.50	0.38	0.57
% Strong disagreement	0.10	0.07	0.11	0.15
Number of Portfolios	559	288	93	178
Number of Loan Officers	176	156	85	133

Notes: This table presents summary statistics of types of questions asked to a generative AI chatbot by loan officers. Based on an analysis of the chatbot's initial response and the officer's subsequent question, I categorized the interactions into three types: information-seeking or neutral questions, weak disagreement, and strong disagreement.

Table A10. Binary Indicator of Whether a Question is Asked to the Chatbot, by Portfolio Gender

Outcome: Question Asked Binary	(1)	(2)	(3)	(4)
	All	Approval by AI (Ex-post Repaid)	Approval by AI (Ex-post Defaulted)	Rejection by AI (Ex-post Defaulted)
Male Names	0.01 (0.02)	0.03 (0.02)	-0.08** (0.04)	0.02 (0.04)
% Portfolios Questioned (Male)	0.24	0.22	0.15	0.38
% Portfolios Questioned (Female)	0.23	0.18	0.23	0.36
Number of Observations	2400	1440	480	480
Number of Loan Officers	240	240	240	240

Notes: The outcome is a binary variable equal to 1 if a question was asked about the portfolio by loan officers, and 0 if no question was asked. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A11. Intensity of Questions Asked to the Chatbot, by Portfolio Gender

Outcome: Intensity of Questions	(1)	(2)	(3)	(4)
	All	Approval by AI (Ex-post Repaid)	Approval by AI (Ex-post Defaulted)	Rejection by AI (Ex-post Defaulted)
Male Names	0.00 (0.03)	0.06 (0.04)	-0.15** (0.06)	-0.01 (0.07)
Intensity of Questions (Male)	0.40	0.36	0.23	0.69
Intensity of Questions (Female)	0.39	0.30	0.38	0.70
Number of Observations	2400	1440	480	480
Number of Loan Officers	240	240	240	240

Notes: The intensity of questions ranges from 0 to 3: 0 indicates no question, 1 indicates an information-seeking or neutral question, 2 indicates weak disagreement, and 3 indicates strong disagreement. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A12. Heterogeneity by Officers' Characteristics: Gender Gap in Loan Approval

	Gender		IAT Score		Experience		Education	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Male	Female	Biased against Women (≥ 0.35)	Others (< 0.35)	Above Median	Below Median	College or Above	Others
Male Names	0.044 (0.028)	0.044 (0.027)	0.048* (0.027)	0.030 (0.032)	0.018 (0.026)	0.073** (0.028)	0.025 (0.025)	0.083*** (0.031)
Female Loan Approval Rate Mean	0.73	0.69	0.70	0.72	0.73	0.68	0.72	0.68
Number of Portfolios	900	900	980	760	940	860	1160	640
Number of Loan Officers	90	90	98	76	94	86	116	64
T-test p-values		1.000		0.674		0.145		0.141

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the heterogeneity in the gender gap in loan approval based on loan officers' characteristics. The IAT score refers to the Gender-Career Implicit Association Test (IAT) score, while Experience denotes the number of years the loan officer has worked in that role. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A13. Heterogeneity by Officers' Characteristics: Debiasing Treatment Effects in Loan Approval

	Gender		IAT Score		Experience		Education	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Male	Female	Biased against Women (≥ 0.35)	Others (< 0.35)	Above Median	Below Median	College or Above	Others
IAT Feedback (T1) x Male Names	-0.029 (0.036)	-0.049 (0.038)	-0.026 (0.039)	-0.040 (0.037)	-0.024 (0.036)	-0.044 (0.038)	-0.034 (0.033)	-0.061 (0.042)
Higher Performance Pay (T2) x Male Names	-0.060 (0.038)	-0.071* (0.038)	-0.078** (0.035)	-0.022 (0.043)	-0.018 (0.035)	-0.119*** (0.040)	-0.033 (0.035)	-0.118*** (0.041)
Combined (T3) x Male Names	-0.033 (0.037)	-0.062 (0.039)	-0.064* (0.038)	-0.017 (0.040)	-0.025 (0.035)	-0.061 (0.042)	-0.037 (0.034)	-0.065 (0.044)
Male Names	0.044 (0.028)	0.044 (0.028)	0.047* (0.026)	0.028 (0.030)	0.018 (0.026)	0.068** (0.027)	0.026 (0.025)	0.080*** (0.030)
T-test p-values (T1)		0.705		0.794		0.697		0.611
T-test p-values (T2)		0.836		0.313		0.061		0.116
T-test p-values (T3)		0.594		0.393		0.502		0.626
Number of Portfolios	3600	3600	3760	3330	4070	3130	4420	2780
Number of Loan Officers	360	360	376	333	407	313	442	278

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents the heterogeneity in changes in loan approval rates and screening time across treatment arms, relative to the control group, based on loan officers' characteristics. The IAT score refers to the Gender-Career Implicit Association Test (IAT) score, while Experience denotes the number of years the loan officer has worked in that role. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A14. Heterogeneity by Officers' Characteristic: AI-assisted Decision and Loan Screening Performance

Loan Officers' Characteristics						
	Gender		IAT Score		Experience	
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Biased against Women (≥ 0.35)	Others (< 0.35)	Above Median	Below Median
Generative AI	0.068*** (0.019)	0.064*** (0.019)	0.063*** (0.017)	0.077*** (0.021)	0.038** (0.016)	0.111*** (0.025)
Non-Generative AI	0.078*** (0.019)	0.100*** (0.019)	0.077*** (0.018)	0.112*** (0.020)	0.074*** (0.017)	0.118*** (0.023)
2nd Control Accuracy Mean	0.544	0.562	0.571	0.530	0.570	0.528
T-test Pvalue (Gen AI)		0.879		0.614		0.013
T-test Pvalue (Non-Gen AI)		0.392		0.187		0.121
Number of Portfolios	3600	3600	3900	3300	4070	3130
Number of Loan Officers	360	360	390	330	407	313
	Education		AI Use Experience Before			
	College or Above					
	College or Above	Others	Yes	No		
Generative AI	0.073*** (0.018)	0.053** (0.021)	0.093*** (0.022)	0.053*** (0.017)		
Non-Generative AI	0.088*** (0.017)	0.086*** (0.022)	0.117*** (0.023)	0.074*** (0.016)		
2nd Control Accuracy Mean	0.553	0.553	0.537	0.562		
T-test Pvalue (Gen AI)		0.456		0.146		
T-test Pvalue (Non-Gen AI)		0.936		0.129		
Number of Portfolios	4420	2780	2260	4940		
Number of Loan Officers	442	278	226	494		

Notes: This table shows the heterogeneity in the improvement in second-round loan screening performance (accuracy) resulting from the use of AI by loan officers' characteristics. Screening performance is a binary variable equal to 1 if an ex-post repaid loan is approved or an ex-post defaulted loan is rejected. It is equal to 0 if an ex-post defaulted loan is approved or an ex-post repaid loan is rejected. The IAT score refers to the Gender-Career Implicit Association Test (IAT) score, while Experience denotes the number of years the loan officer has worked in that role. AI use experience is a binary variable equal to 1 if a loan officer has any prior experience using generative AI tools such as ChatGPT, and 0 otherwise. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A15. Heterogeneity: Selective Follow of AI Recommendations for Ex-post Defaulted Loan Approvals

Loan Officers' Characteristics						
	Gender		IAT Score		Experience	
	(1)	(2)	(3)	(4)	(5)	(6)
	Male	Female	Biased against Women (≥ 0.35)	Others (< 0.35)	Above Median	Below Median
Gen AI Binary x Male Names	0.117* (0.061)	0.108* (0.058)	0.083 (0.061)	0.143** (0.061)	0.112** (0.055)	0.122* (0.065)
Gen AI Binary	-0.125** (0.048)	0.000 (0.046)	-0.001 (0.047)	-0.126** (0.049)	-0.091** (0.044)	-0.032 (0.051)
Male Names	0.008 (0.039)	-0.058 (0.045)	-0.000 (0.044)	-0.051 (0.041)	-0.033 (0.037)	-0.015 (0.047)
T-test p-value (Gen AI Binary x Male Names)		0.921		0.489		0.906
Number of Portfolios	480	480	510	450	522	438
Number of Loan Officers	240	240	255	225	261	219

	Education		AI Use Experience Before	
	College or Above	Others	Yes	No
Gen AI Binary x Male Names	0.199*** (0.056)	-0.003 (0.062)	0.239*** (0.086)	0.048 (0.048)
Gen AI Binary	-0.140*** (0.043)	0.050 (0.052)	-0.142** (0.066)	-0.022 (0.039)
Male Names	-0.067* (0.038)	0.045 (0.046)	-0.085 (0.060)	0.006 (0.035)
T-test p-value (Gen AI Binary x Male Names)		0.016		0.053
Number of Portfolios	588	372	284	676
Number of Loan Officers	294	186	142	338

Notes: This table shows the heterogeneity in the loan officers' follow rate for AI recommendations in the second round by loan officers' characteristics. The AI is programmed to make three types of recommendations: approve ex-post repaid loans, approve ex-post defaulted loans, and reject ex-post defaulted loans. The outcome is a binary follow-rate variable, equal to 1 if the officer follows the AI's recommendation and 0 otherwise. The IAT score refers to the Gender-Career Implicit Association Test (IAT) score, while Experience denotes the number of years the loan officer has worked in that role. AI use experience is a binary variable equal to 1 if a loan officer has any prior experience using generative AI tools such as ChatGPT, and 0 otherwise. Standard errors, clustered at the loan officer level, are reported in parentheses. Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Appendix B: Process for Selecting Experimental Loan Applications

B.1. Procedure for Selecting the 50 Loans Used in the Experiment

The 50 loans used in this experiment were selected from a pool of 200 previously approved applications provided by the Family Development Foundation. From this initial set, 49 loans were excluded because their tenures had not yet concluded, leaving 151 eligible applications. Based on the power analysis, the maximum number of loans that could be used to achieve sufficient statistical power was 50. To arrive at the final sample of 50 loans, a multi-stage filtering procedure was implemented. This procedure was designed to mitigate potential statistical discrimination and to isolate "borderline" cases where loan officer discretion is most pronounced.

The first filter addressed industry-level gender representation. To reduce the potential impact of statistical bias, the sample was restricted to industries where each gender constituted at least 30% of borrowers. This step ensured that an applicant's gender was not correlated with industry-specific risk factors. The qualifying industries included cultivation, supermarkets, grocery stores, clothing shops, fish sale shops, poultry, and livestock businesses.

The second filter focused on identifying borderline applications where loan officers have greater discretion and gender bias is more likely to emerge (Costello et al., 2020; Agarwal et al., 2024; FDIC, 2024). In the Egyptian credit scoring system, scores are categorized into five tiers: Excellent (above 750), Very Good (701–750), Satisfactory (626–700), Not Satisfactory (520–625), and Bad (below 520). Applications with clear-cut approval or rejection signals were excluded, specifically those with "Excellent" scores (above 750) or "Bad" scores (below 520). The selection was narrowed to loans with scores between 520 and 750, as well as those with no credit score, which is common for entrepreneurs in the informal sector. After applying both the industry and credit score filters, 68 candidate loans remained.

From these 68 candidates, the final 50 loans were selected using a machine learning (ML) methodology designed to identify loans with the most predictable outcomes based on application data. The objective was not to develop a high-accuracy prediction model but rather to filter out loans whose ex-post outcomes appeared random or noisy. To reduce the over-fitting issue from a small sample size, an ensemble of seven ML algorithms was employed: Lasso, Random Forest, Support Vector Machines (SVM), Neural Networks, Gradient Boosting, AdaBoost, and HistGradientBoost. These models were selected for their predictive performance in related literature (Cheraghali and Molnár, 2024) and their widespread use in economics and applied ML. The selection process involved 100 independent iterations. In each iteration, the full 151-loan dataset was randomly split into a 70% training set and a 30% test set. For each loan and each of the seven models, a prediction was deemed accurate if its predicted probability of repayment (threshold > 0.5) matched

its ex-post outcome (repaid or defaulted). A final "predictability benchmark" score was then calculated for each of the 68 candidate loans by averaging the accuracy scores from its four best-performing models across all iterations in which it appeared in a test set.

Finally, the 50-loan sample was constructed by stratifying based on gender and loan outcome, using the predictability benchmark to guide selection. The target sample composition was 30 repaid loans (15 male, 15 female) and 20 defaulted loans (10 male, 10 female). The pool of 68 candidates contained 17 male-repaid, 28 female-repaid, 11 male-defaulted, and 12 female-defaulted loans. To meet the target strata, loans with the lowest predictability scores within each category were removed. This entailed excluding the 2 lowest-scoring male-repaid, 13 lowest-scoring female-repaid, 1 lowest-scoring male-defaulted, and 2 lowest-scoring female-defaulted loans.

This ML-based selection method has limitations, including a non-standard cross-validation design and a relatively low average Area Under the Curve (AUC) of approximately 0.60 across iterations, reflecting the small sample size. However, a post-hoc validation exercise suggests the method's utility. Using an expanded dataset of 500 additional loans after the experiment, the selection procedure was repeated 1,000 times with a Gradient Boosting Machine (GBM) model introduced in Section 4.2.4 with AUC 0.88. It is trained on 542 loans and tested on the original 68. This procedure re-selected an average of 44 of the original 50 loans, yielding an 88% overlap. This result indicates that, despite its limitations, the selection process likely performed substantially better than random sampling by successfully filtering out loans with highly unpredictable outcomes.

B.2. Procedure for Grouping the 50 Loans into Bundles

For the experiment, the 50 selected loans were organized into five bundles, each containing 10 loans. A simpler approach, such as randomly assigning 10 individual loans to each officer from the total pool of 50, was intentionally avoided. Under such a design, two officers evaluating the same loan application (where only the applicant's gender was experimentally altered) would make decisions within the context of two different sets of nine other loans. The composition of an evaluation batch can influence a loan officer's judgment. For example, an officer's assessment of a specific borderline application might become stricter if it is evaluated alongside exceptionally strong applications, or more lenient if it is surrounded by particularly weak ones. These spillover effects could introduce systematic noise, making it difficult to isolate the causal impact of applicant gender. By creating fixed bundles, this procedure ensures that when different loan officers evaluate a given application, they do so within an identical context of the other nine loans in the bundle.

To balance the overall difficulty level across the five bundles, I used the predictability benchmark derived

from the machine learning (ML) model described in Appendix B.1. This benchmark captures how well each loan’s repayment outcome could be predicted from its application data. Within each subgroup defined by applicant gender and ex-post repayment status (i.e., male-repaid, female-repaid, male-defaulted, and female-defaulted), I ordered the loans by their predictability scores. I assigned the loans to bundles using a symmetric "zigzag" pattern to ensure an even distribution of predictability across the five bundles. For instance, among the 15 ex-post male-name repaid loans, I assigned the most predictable loan to Bundle 1, the second to Bundle 2, the third to Bundle 3, and so on until Bundle 5. Then, I reversed the direction: the sixth most predictable loan went into Bundle 5, the seventh into Bundle 4, and so on back to Bundle 1. This zigzag pattern continued until all 15 loans were assigned. I applied the same procedure separately to the 15 ex-post female repaid loans, 10 male defaulted loans, and 10 female defaulted loans.

Finally, I created five additional “clone” bundles (Bundles 6 to 10) that are exact clones of Bundles 1 to 5 in terms of loan content and order, but with first names switched to indicate the opposite gender. For example, if the first three loans in Bundle 1 featured applicants with genders female, male, and female, then the corresponding loans in Bundle 6 would have the same loan details and ordering, but with the genders switched to male, female, and male, respectively. To implement the gender manipulation, I removed the gender variable from the loan application and replaced the applicant’s real first name with a common male or female name to signal gender. The replacement names were selected from the top 10 most frequently appearing male and female first names in the loan portfolios provided by the Family Development Foundation. The male names included Muhammad, Ahmed, Abdul, Mostafa, Mahmoud, Hussein, Ali, Hassan, Jaber, and Ashraf; the female names included Fatima, Karima, Iman, Shadia, Soheir, Hamida, Hana, Heba, Marwa, and Noha. This allows for a clean experimental comparison: by comparing loan officers assigned to Bundle 1 with those assigned to Bundle 6, I can examine how the same portfolio of loan applications is evaluated differently when only the applicant’s gender is altered.

Appendix C: Prompts used in Gpt-4o

The prompt used to instruct GPT-4o to predict the defaulted loans for the 100 loans in the experiment is as follows. This prompt was used to evaluate the accuracy of GPT-4o in predicting the default outcomes.

PROMPT A:

Task: Predict whether a loan is likely to have a late repayment of 90 days or more, which is regarded as a default loan. The profiles consist of small business owners in Egypt. In general, the probability of having 90 days or more late repayment is 40%. Additionally, the probability is the same regardless of gender.

Instructions: Based on the portfolio profile, predict if a late repayment of 90 days or more is likely. Respond with: '1' if late repayment of 90 days or more is likely. '0' if late repayment of 90 days or more is unlikely.

Profile of a client:

Name: NAME

Age: AGE

The number of guarantors who are employees: GURAN_EMPLOYEE

The number of guarantors who are not employees: GURAN_PERSONAL

The client's average monthly income in EGP : MONTHLY_INCOME

Whether the business is registered with the government (1=Yes, 0=No): REGISTERED

The year the business was established: BIZ_START_YEAR

The region where the business is located in Egypt: REGION

The number of employees in the business: NUM_EMPLOYEE

The industry in which the client's business operates: INDUSTRY

The amount of the loan granted: LOAN_SIZE

The loan's tenure in month: TENURE

The current interest rate on the loan: IR_CURRENT

Number of previous loans the client has taken: NUM_PRE_LOANS

The number of delayed days on the previous loan: DELAYED_DAYS_PAST

The amount of the loan granted on the previous loan: LOAN_SIZE_PAST

The interest rate on the previous loan: IR_PAST

The client's credit score: I_SCORE

Next, in the main experiment, I provided GPT-4o with a portfolio decision (either approval or rejection) and asked GPT-4o to generate two reasons supporting the given decision, as shown in Figures A4 and A5. The prompt used for this task is as follows:

PROMPT B:

CHATBOT PREFIX = Below is the loan portfolio of the small business owner in Egypt. The loan decision is made based on the likelihood of late repayment of 90 days or more. For example, if late repayment of 90 days or more is unlikely, it is approved and rejected otherwise.

PROMPT TEMPLATE = Profile of a client

Name: NAME

Age: AGE

The number of guarantors who are employees: GURAN_EMPLOYEE

The number of guarantors who are not employees: GURAN_PERSONAL

The client's average monthly income in EGP : MONTHLY_INCOME

Whether the business is registered with the government (1=Yes, 0=No): REGISTERED

The year the business was established: BIZ_START_YEAR

The region where the business is located in Egypt: REGION

The number of employees in the business: NUM_EMPLOYEE

The industry in which the client's business operates: INDUSTRY

The amount of the loan granted: LOAN_SIZE

The loan's tenure in month: TENURE

The current interest rate on the loan: IR_CURRENT

Number of previous loans the client has taken: NUM_PRE_LOANS

The number of delayed days on the previous loan: DELAYED_DAYS_PAST

The amount of the loan granted on the previous loan: LOAN_SIZE_PAST

The interest rate on the previous loan: IR_PAST

The client's credit score: I_SCORE

LOAN DECISION: APPROVE_DECISION

REASON_Q = What are the reasons for the loan decision? Answer with two reasons.

Appendix D: Survey Questions

This section explains the baseline survey questions presented in Table 1. The "Loan Officers Skill Binary", "Cognitive Test", "Digit Span Recall", "Psychometric", and "Risk Attitude" questions are adopted from Bryan et al. (2024). The Social Desirability Test questions are from Hays et al. (1989a,b), and the Explicit Bias Questions are taken from the World Values Survey (Haerpfer et al., 2022). The format and instructions for the Gender-Career Implicit Association Test (IAT) are based on Greenwald et al. (1998, 2003); Lane et al. (2007), and <https://implicit.harvard.edu/implicit/>.

Loan Officers Skill Binary

There are two types of loan officers. Which characteristics do you think are similar to you?

- **PERSON A:** Person A has good communication skills; he is patient and a good listener and also has the ability to convince people, motivate them and give them valuable feedback.
- **PERSON B:** Person B is firm in his dealings and knows how to say no when required. He has good analytical skills along with excellent supervisory and management skills.

Cognitive Test

Now I'm going to ask you a series of intelligence and general knowledge questions.

- Q1. A bat and a ball cost \$11 in total. The bat costs 1 dollars more than the ball. How much does the ball cost?
- Q2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
- Q3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake from the beginning?

Digit Span Recall

Now we're going to play a game. A series of digits will appear for a few seconds. Try to remember the digits and repeat them back in order. After each successful turn an extra digit will be added.

Social Desirability Test

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- I am always courteous even to people who are disagreeable.
- There have been occasions when I took advantage of someone.
- I sometimes try to get even rather than forgive and forget.
- I sometimes feel resentful when I don't get my way.
- No matter who I'm talking to, I'm always a good listener.

Psychometric Questions

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- I tend to act first and worry about the consequences later.
- I can think of several solutions to any problem.
- When I make decisions I usually go with my first gut feeling.
- I always get things done ahead of time.
- I would work seven days a week if I could.
- I spend a lot of time planning for my future.
- I prefer to have a flexible schedule - I don't like being tied down.

Explicit Bias Questions

Now I will read some phrases to you, and I want you to rate your level of agreement for each one. Please select one of the following options for each statement: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree.

- When a mother works for pay, the children suffer.

- On the whole, men make better political leaders than women do.
- A university education is more important for a boy than for a girl.
- On the whole, men make better business executives than women do.
- If a woman earns more money than her husband, it's almost certain to cause problems.
- When jobs are scarce, men should have more right to a job than women.

Risk Attitude

- R1: How do you see yourself? Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks? Please select one number on the scale, where 0 equals "unwilling to take risks" and 10 equals: "fully prepared to take risks".
- R2: Imagine you have EGP 10,000 to invest. You may choose to "invest" all or a portion of your money. There is a 50% chance to double the amount of money you choose to "invest". It is equally possible that you could lose half the investment. For example, if you choose to invest EGP 1,000, there is a 50-50 chance of getting EGP 2,000 or EGP 500. How much of the EGP 10,000 do you invest?

IAT Introduction

You will be presented with a series of nouns that are typically associated with Career or Family. You must state if that word is most related to career or family. To select "Career", press E on your keyboard. To select "Family", press I on your keyboard. If you select the incorrect answer, we will let you know, and you should then select the correct answer.

Answer as quickly as possible. When you are ready, go to the next field. There, you will either press the spacebar or tap the box at the bottom to start.

IAT Feedback: Control and Higher Performance Pay Groups

Thanks for your participation. Let's go to the next page.

IAT Feedback: IAT Feedback and Combined Intervention Groups

Thanks for your participation. Your Gender-Career IAT score is {IAT_score_final}. Here's how to understand your results:

If your score is positive, it indicates a stronger subconscious association between "male" and "career" or "female" and "family." Conversely, if your score is negative, it signifies a stronger subconscious association between "male" and "family" or "female" and "career." Here's how to interpret the magnitude of your score:

- No association if the score is between -0.15 and 0.15 ,
- Slight association for values between $|0.15|$ and $|0.35|$,
- Moderate association between $|0.35|$ and $|0.65|$,
- Strong association for scores higher than $|0.65|$.

We want to iterate that this test reveals implicit attitudes and not behaviors. Our attitudes may derive from the cultural and social context where we live. An enormous body of literature confirms that we all have biases—some explicit, many implicit. However, it is important to avoid our implicit biases or stereotypes related to a specific group from systematically influencing our behavior toward clients. Acknowledging and understanding our biases and those of our colleagues can help minimize the influence they have on our daily interaction with clients.

Appendix E: Theoretical Model Derivations

This appendix provides the detailed derivations for the theoretical model presented in Section 2.

E.1. Bayesian Updating of Posterior Beliefs

Before exerting effort—i.e., prior to receiving any signal—the loan officer holds a prior belief about the revenue r of an applicant, modeled as:

$$r \sim \mathcal{N}(m_{G_j}, \sigma_r^2)$$

where m_{G_j} is the prior mean revenue for an applicant of gender G_j , and σ_r^2 is the prior variance.

After exerting effort e_j , the officer observes a noisy signal:

$$s = r + \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}\left(0, \frac{1}{e_j}\right)$$

Using Bayes' rule, the officer updates their belief about the revenue r . The posterior distribution is:

$$r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2)$$

with posterior mean and variance given by:

$$\hat{m}_{G_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2}, \quad \hat{\sigma}^2 = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

These expressions reflect precision-weighted updating, where the prior and signal are combined based on the inverse of their variances (i.e., their precision).

- **Posterior Mean (\hat{m}):** A weighted average of the prior mean and the observed signal, with weights proportional to their precision:

$$\hat{m}(s, e_j, G_j) = \frac{\left(\frac{1}{\sigma_r^2}\right) m_{G_j} + e_j s}{\frac{1}{\sigma_r^2} + e_j} = \frac{m_{G_j} + e_j \sigma_r^2 s}{1 + e_j \sigma_r^2}$$

- **Posterior Variance ($\hat{\sigma}^2$):** The inverse of the total precision:

$$\hat{\sigma}^2(e_j) = \frac{1}{\frac{1}{\sigma_r^2} + e_j} = \frac{\sigma_r^2}{1 + e_j \sigma_r^2}$$

E.2. Derivation of Decision Thresholds

Derivation of the Approval Probability Threshold (k_G^*)

The officer approves an application if the expected utility of approving, $E[U(\text{Approve})]$, exceeds that of rejecting, $E[U(\text{Reject})]$. The threshold k_G^* is the posterior probability of repayment, $p(s, e_j, G_j)$, where the officer is exactly indifferent. I find this by setting the expected utilities equal:

$$E[U(\text{Approve})] = E[U(\text{Reject})]$$

Let $p \equiv p(s, G_j)$. The expected utilities for each action are:

$$E[U(\text{Approve})] = p \cdot (+\pi\ell_2) + (1 - p) \cdot (-\pi\ell_1) + d_j(G_j)$$

$$E[U(\text{Reject})] = p \cdot (-\pi\ell_2) + (1 - p) \cdot (+\pi\ell_1)$$

Setting these equal and solving for p yields the threshold k_G^* :

$$\begin{aligned} \pi p \ell_2 - \pi(1 - p)\ell_1 + d_j(G_j) &= -\pi p \ell_2 + \pi(1 - p)\ell_1 \\ k_G^* &= \frac{2\pi\ell_1 - d_j(G_j)}{2\pi(\ell_1 + \ell_2)} \end{aligned}$$

Substituting $d_j(M) = \delta_M$ and $d_j(F) = -\delta_F$, I obtain:

$$k_M^* = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1 + \ell_2)} \leq k_F^* = \frac{2\pi\ell_1 + \delta_F}{2\pi(\ell_1 + \ell_2)}$$

where $0 \leq \delta_M \leq 2\pi\ell_1$ and $0 \leq \delta_F \leq 2\pi\ell_2$.

Derivation of the Signal Threshold (s_G^*)

The officer's belief about the loan being repaid is captured by $p(s, e, g)$, defined as the posterior the loan's revenue $r \geq 1$, given the observed signal s_j by effort e_j and the applicant's gender G_j .

$$p(s_j, e_j, G_j) = P(r \geq 1 \mid s_j, e_j, G_j) = 1 - P(r < 1 \mid s_j, e_j, G_j)$$

Since $r \mid s, e_j, G_j \sim \mathcal{N}(\hat{m}_{G_j}, \hat{\sigma}^2)$, this probability can be expressed using the standard normal CDF $\Phi(\cdot)$:

$$p(s, e_j, G_j) = 1 - \Phi\left(\frac{1 - \hat{m}(s, e_j, G_j)}{\hat{\sigma}(e_j)}\right) = \Phi\left(\frac{\hat{m}(s, e_j, G_j) - 1}{\hat{\sigma}(e_j)}\right) = \Phi\left(\frac{(m_{G_j} - 1) + e_j \sigma_r^2 (s - 1)}{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)}}\right)$$

The signal threshold, s_G^* , is the signal value that yields a posterior probability equal to k_G^* . I solve for s_G^* by setting $p(s_G^*) = k_G^*$:

$$\begin{aligned} \Phi\left(\frac{(m_{G_j} - 1) + e_j \sigma_r^2 (s_G^* - 1)}{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)}}\right) &= k_G^* \\ \frac{(m_{G_j} - 1) + e_j \sigma_r^2 (s_G^* - 1)}{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)}} &= \Phi^{-1}(k_G^*) \\ s_G^* &= 1 + \frac{\sqrt{\sigma_r^2 (1 + e_j \sigma_r^2)} \cdot \Phi^{-1}(k_G^*) - (m_{G_j} - 1)}{e_j \sigma_r^2} \end{aligned}$$

Then, the difference in approval thresholds between male (M) and female (F) applicants is:

$$\begin{aligned} s_M^* - s_F^* &= \frac{\sqrt{\sigma_r^2 (1 + e_M^* \sigma_r^2)} \cdot \Phi^{-1}(k_M^*) - (m_M - 1)}{e_M^* \sigma_r^2} - \frac{\sqrt{\sigma_r^2 (1 + e_F^* \sigma_r^2)} \cdot \Phi^{-1}(k_F^*) - (m_F - 1)}{e_F^* \sigma_r^2} \\ &= \underbrace{\left(\frac{\sqrt{(1 + e_M^* \sigma_r^2)} \cdot \Phi^{-1}(k_M^*)}{e_M^* \sigma_r} - \frac{\sqrt{(1 + e_F^* \sigma_r^2)} \cdot \Phi^{-1}(k_F^*)}{e_F^* \sigma_r} \right)}_{\text{Taste-Based Bias}} + \underbrace{\left(\frac{1 - m_M}{e_M^* \sigma_r^2} - \frac{1 - m_F}{e_F^* \sigma_r^2} \right)}_{\text{Prior Belief}} \end{aligned}$$

E.3. Derivation of the First-Order Condition

The officer chooses effort e to maximize expected utility (with convex cost ce^2):

$$U(e; G, \pi, d) = \int_{-\infty}^{s_G^*(e, \pi, d)} V_R(s, e, G, \pi) g(s | e, G) ds + \int_{s_G^*(e, \pi, d)}^{\infty} V_A(s, e, G, \pi, d) g(s | e, G) ds - ce^2,$$

where $V_R(s, e, G, \pi) = \pi[\ell_1 - (\ell_1 + \ell_2)p(s, e, G)]$, $V_A(s, e, G, \pi, d) = \pi[-\ell_1 + (\ell_1 + \ell_2)p(s, e, G)] + d(G)$, and $g(\cdot | e, G)$ is the pdf of the signal.

Differentiating with respect to e using Leibniz' rule yields

$$\frac{dU}{de} = \int_{-\infty}^{s^*} \partial_e(V_R g) ds + \int_{s^*}^{\infty} \partial_e(V_A g) ds - 2ce,$$

since boundary terms vanish by the cutoff condition $V_A(s^*) = V_R(s^*)$. Applying the product rule $\partial_e(Vg) =$

$V_e g + V g_e$, and using $V_{R,e} = -\pi(\ell_1 + \ell_2)p_e$, $V_{A,e} = \pi(\ell_1 + \ell_2)p_e$, gives

$$\frac{dU}{de} = \int_{s^*}^{\infty} (\pi(\ell_1 + \ell_2)p_e g + V_A g_e) ds + \int_{-\infty}^{s^*} (-\pi(\ell_1 + \ell_2)p_e g + V_R g_e) ds - 2ce.$$

Grouping terms, the first-order condition is

$$\boxed{\frac{dU}{de} = \pi A(e; k_G^*(\pi, d(G))) + d(G) B(e; k_G^*(\pi, d(G))) - 2ce.}$$

where $A(e; k) = \int_{s^*}^{\infty} ((\ell_1 + \ell_2)p_e g + [-\ell_1 + (\ell_1 + \ell_2)p] g_e) ds + \int_{-\infty}^{s^*} ((\ell_1 + \ell_2)p_e g + [\ell_1 - (\ell_1 + \ell_2)p] g_e) ds$,

$$B(e; k) = \int_{s^*}^{\infty} g_e(s | e, G) ds, \quad g_e(s | e, G) = \frac{\partial g(s | e, G)}{\partial e} \quad \text{and} \quad p_e(s | e, G) = \frac{\partial p(s | e, G)}{\partial e}.$$

Here, $A(e; k)$ is the monetary channel (stakes \times informativeness), strictly positive under the standard regularity assumption that higher effort increases signal precision, while $B(e; k)$ captures how effort shifts the approval region above s^* , scaling $d(G)$. The sign of B_k is not definitive, as it depends on how effort shifts the signal distribution around the threshold s^* .

E.4. Proof of Proposition 1

Proposition 1. *If effort e is sufficiently large, the presence of biased prior belief ($m_M > m_F$) or taste-based bias ($\delta_M > 0$ or $\delta_F > 0$) leads to a lower approval threshold for male applicants ($s_M^* < s_F^*$).*

The difference in approval thresholds between applicant genders can be decomposed into two components:

$$\begin{aligned} s_M^* - s_F^* = & \underbrace{\left(\frac{\sqrt{1 + e_M^* \sigma_r^2}}{e_M^* \sigma_r} \cdot \Phi^{-1}(k_M^*) - \frac{\sqrt{1 + e_F^* \sigma_r^2}}{e_F^* \sigma_r} \cdot \Phi^{-1}(k_F^*) \right)}_{\text{Taste-Based Bias Effect}} \\ & + \underbrace{\left(\frac{1 - m_M}{\sigma_r^2 e_M^*} - \frac{1 - m_F}{\sigma_r^2 e_F^*} \right)}_{\text{Prior Belief Effect}}. \end{aligned}$$

- **Prior-belief-driven gender gap:** For sufficiently large e , both $\frac{\sqrt{1 + e \sigma_r^2}}{e \sigma_r}$ and $\frac{1}{\sigma_r^2 e}$ are decreasing and convex in e . As a result, differences between e_M^* and e_F^* have only a negligible effect. Consequently, the taste-based bias term is approximately zero when $k_M^* = k_F^*$. In contrast, the prior-belief effect remains strictly negative when $m_M > m_F > 1$, implying that $s_M^* < s_F^*$.
- **Taste-based-bias-driven gender gap:** When $m_M = m_F$ and e is sufficiently large, the prior-belief

effect is negligible because $\frac{1}{\sigma_e^2}$ is decreasing and convex in e . Taste-based bias, however, implies $\Phi^{-1}(k_M^*) < \Phi^{-1}(k_F^*)$. Since the effect of differences in e_M^* and e_F^* remain negligible, the taste-based bias term (either $\delta_M > 0$ or $-\delta_F < 0$) is negative for sufficiently large effort, yielding $s_M^* < s_F^*$.

E.5. Proof of Proposition 2

Proposition 2. *For a sufficiently large increase in performance pay (π), the officer's optimal screening effort rises.*

To study how performance pay π affects optimal effort, define the first-order condition

$$F(e, \pi; d) \equiv \pi A(e; k_G^*(\pi, d)) + d(G) B(e; k_G^*(\pi, d)) - 2ce,$$

where $k_G^*(\pi, d) = \frac{2\pi\ell_1 - d(G)}{2\pi(\ell_1 + \ell_2)}$ and $\frac{\partial k_G^*}{\partial \pi} = \frac{d(G)}{2\pi^2(\ell_1 + \ell_2)} \geq 0$.

The optimal effort e_G^* is characterized by $F(e_G^*, \pi; d) = 0$. By the implicit function theorem,

$$\frac{de_G^*}{d\pi} = -\frac{F_\pi}{F_e} \Big|_{e=e_G^*}.$$

If the second-order condition $F_e < 0$ holds, the sign of $\frac{de_G^*}{d\pi}$ is determined by the sign of F_π :

$$F_\pi = A(e_G^*; k_G^*) + \frac{d(G)}{2\pi(\ell_1 + \ell_2)} A_k(e_G^*; k_G^*) + \frac{d(G)^2}{2\pi^2(\ell_1 + \ell_2)} B_k(e_G^*; k_G^*),$$

where subscripts denote partial derivatives with respect to k .

- $A(e; k) > 0$ is the marginal monetary return to effort (via repayment payoffs).
- A_k measures how the monetary return changes as the cutoff shifts; its sign is ambiguous because a tighter cutoff both reduces the mass affected by effort and increases the sensitivity of marginal loans.
- $B_k \leq 0$ under standard approval monotonicity, since tightening the cutoff reduces approval probability.

As $\pi \rightarrow \infty$, $k_G^*(\pi, d)$ converges to $\bar{k} = \ell_1/(\ell_1 + \ell_2)$, and $A(e_G^*; k_G^*)$ converges to a positive finite limit, while A_k and B_k remain bounded. Meanwhile, both $\frac{d(G)}{2\pi(\ell_1 + \ell_2)}$ and $\frac{d(G)^2}{2\pi^2(\ell_1 + \ell_2)}$ converge to zero. Therefore,

$$\lim_{\pi \rightarrow \infty} F_\pi = A(e_G^*; k_G^*) > 0.$$

It follows that for sufficiently large π ,

$$\frac{de_G^*}{d\pi} = -\frac{F_\pi}{F_e} > 0.$$

E.6. Proof of Proposition 3

Proposition 3. *If either δ_M or m_M is sufficiently large, higher performance pay raises the approval threshold for male applicants (s_M^*).*

The male approval threshold is:

$$s_M^* = 1 + \underbrace{\frac{\sqrt{\sigma_r^2(1 + e_M^* \sigma_r^2)} \cdot \Phi^{-1}(k_M^*)}{e_M^* \sigma_r^2}}_{\text{Taste-Based Bias Effect}} - \underbrace{\frac{(m_M - 1)}{e_M^* \sigma_r^2}}_{\text{Prior Belief Effect}}, \quad k_M^* = k_M^*(\pi, \delta_M) = \frac{2\pi\ell_1 - \delta_M}{2\pi(\ell_1 + \ell_2)}.$$

Let $a = \sigma_r^2$, $e = e_M^*(\pi)$, $k = k_M^*(\pi, \delta_M)$, $z = \Phi^{-1}(k)$, and

$$f(e) = \frac{\sqrt{1+ea}}{e\sqrt{a}}, \quad f'(e) = -\frac{ea+2}{2e^2\sqrt{a}\sqrt{1+ea}}, \quad \frac{d}{dk}\Phi^{-1}(k) = \frac{1}{\phi(\Phi^{-1}(k))}.$$

By the chain rule,

$$\frac{ds_M^*}{d\pi} = \left[z f'(e) + \frac{m_M - 1}{e^2 a} \right] \frac{de_M^*}{d\pi} + \frac{f(e)}{\phi(z)} \frac{\partial k_M^*}{\partial \pi}, \quad \frac{\partial k_M^*}{\partial \pi} = \frac{\delta_M}{2\pi^2(\ell_1 + \ell_2)}.$$

By Proposition 2, $\frac{de_M^*}{d\pi} > 0$, and assume $m_M > 1$. Then, $\frac{ds_M^*}{d\pi} > 0$ under either of the following conditions²⁹:

1. Strong taste based bias favoring men: If

$$\delta_M \geq \pi(\ell_1 - \ell_2) \quad \text{with} \quad \ell_1 > \ell_2,$$

then $k_M^*(\pi, \delta_M) \leq \frac{1}{2}$, so $z = \Phi^{-1}(k_M^*) \leq 0$. In this case,

$$\frac{ds_M^*}{d\pi} = \left[z f'(e_M^*) + \frac{m_M - 1}{e_M^{*2} \sigma_r^2} \right] \frac{de_M^*}{d\pi} + \frac{f(e_M^*)}{\phi(z)} \frac{\partial k_M^*}{\partial \pi}, \quad f'(e) = -\frac{e\sigma_r^2 + 2}{2e^2\sigma_r\sqrt{1+e\sigma_r^2}} < 0, \quad \frac{\partial k_M^*}{\partial \pi} = \frac{\delta_M}{2\pi^2(\ell_1 + \ell_2)} > 0.$$

Since $z \leq 0$ and $f'(e_M^*) < 0$, I have $z f'(e_M^*) \geq 0$. With $m_M > 1$, it follows that $\frac{m_M - 1}{e_M^{*2} \sigma_r^2} > 0$. By

²⁹A sufficiently large bias is required because with $\ell_1 > \ell_2$, the sign of z depends on the bias magnitude. On the other hand, when $\ell_1 < \ell_2$, z is always negative, so even a small bias ensures $\frac{ds_M^*}{d\pi} > 0$

Proposition 2, $\frac{de_M^*}{d\pi} > 0$. Moreover, $\frac{f(e_M^*)}{\phi(z)} > 0$ and, as $\delta_M > 0$, $\frac{\partial k_M^*}{\partial \pi} > 0$.

Therefore, $\frac{ds_M^*}{d\pi} > 0$.

2. Strong prior belief in male borrowers' repayment likelihood : If

$$m_M > 1 + z \frac{\sigma_r (e_M^* \sigma_r^2 + 2)}{2 \sqrt{1 + e_M^* \sigma_r^2}},$$

then $z f'(e_M^*) + \frac{m_M - 1}{e_M^{*2} \sigma_r^2} > 0$, hence $\frac{ds_M^*}{d\pi} > 0$ *regardless of whether* $\delta_M > 0$ or $\delta_M = 0$: $\frac{f(e_M^*)}{\phi(z)} \frac{\partial k_M^*}{\partial \pi}$ is strictly positive if $\delta_M > 0$ and equals zero if $\delta_M = 0$.

Therefore, $\frac{ds_M^*}{d\pi} > 0$.