NBER WORKING PAPER SERIES

HOW GOOD IS AI AT TWISTING ARMS? EXPERIMENTS IN DEBT COLLECTION

James J. Choi
Dong Huang
Zhishu Yang
Qi Zhang

Working Paper 33669
http://www.nber.org/papers/w33669

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2025

This paper was formerly titled "Better than Human? Experiments with AI Debt Collectors." This work was supported by the National Natural Science Foundation of China [Grant 72071117]. We thank Anthony DeFusco, Zhao Jin, Sung Kwan Lee, Jiacheng Liu, Xueming Luo, Jinfei Sheng, Haokun Sun, Yijun Zhou, and seminar participants at the ABFER-JFDS Conference, AFA Annual Meeting, China Financial Research Conference, China International Conference in Finance, Columbia & RFS AI in Finance Conference, FMA Annual Meeting, SFS Cavalcade Asia-Pacific, University of Alberta, University of Chicago, Emory, and Yale for helpful discussions and comments. We thank Manlin Sun and Hongyu Wu for excellent research assistance. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Good is AI at Twisting Arms? Experiments in Debt Collection
James J. Choi, Dong Huang, Zhishu Yang, and Qi Zhang
NBER Working Paper No. 33669
April 2025
JEL No. D14, G4, G51, J24

ABSTRACT

How good is AI at persuading humans to perform costly actions? We study calls made to get delinquent consumer borrowers to repay. Regression discontinuity and a randomized experiment reveal that AI is substantially less effective than human callers. Replacing AI with humans six days into delinquency closes much of the gap. But borrowers initially contacted by AI have repaid 1% less of the initial late payment one year later and are more likely to miss subsequent payments than borrowers who were always called by humans. AI's lesser ability to extract promises that feel binding may contribute to the performance gap.

James J. Choi
Yale School of Management
165 Whitney Avenue
P.O. Box 208200
New Haven, CT 06520-8200
and NBER
james.choi@yale.edu

Dong Huang
Yale School of Management
165 Whitney Ave.
New Haven, CT 06511
USA
dong.huang@yale.edu

Zhishu Yang
School of Economics & Management
Tsinghua University
Beijing, 100084
China
yangzhsh@sem.tsinghua.edu.cn

Qi Zhang
Shanghai Jiaotong University
800 Dongchuan RD. Minhang District,
Shanghai,
China
zhang.qi@sjtu.edu.cn

A data appendix is available at http://www.nber.org/data-appendix/w33669

Rapid progress in artificial intelligence (AI) has revived the long-standing debate on the extent to which new technologies will eliminate human jobs.[1] In this paper, we study the effectiveness of AI in a different sort of task than has previously been studied: persuading a human to perform a personally costly action. Many service and managerial jobs require performing this type of task—for example, coaxing a colleague to exert extra uncompensated effort for the good of his team (e.g., serve on a university committee), inducing a customer to make a sacrifice like switching airplane seats so that a family can sit together, or asking somebody to honestly report the details of an accident for insurance claim adjustment purposes.

We find that AI is substantially less effective at "twisting arms." Exposure to AI persuasion also reduces subsequent cooperation with the AI's company. Thus, this sort of influencing may remain an important area of comparative advantage for humans in a future where AI is ubiquitous.

The specific task we study is persuading delinquent consumer borrowers to repay their debt. Contact from a debt collector is a common experience; in 2022, 26% of U.S. adults with a credit bureau record had debt in collections.[2] The job of a debt collector is non-routine, requires social interaction, and is aided by emotional intelligence. In addition, repaying one's debts is usually seen as a moral obligation (Guiso, Sapienza, and Zingales, 2013; Bursztyn et al., 2019), which may cause AI to be less effective than humans at eliciting repayment, since being observed committing a moral transgression by another person is more aversive than being observed by a machine (LaMothe and Bobek, 2020; Cohn et al., 2022; Kim et al., 2023). The moral aspect of the interaction is not unusual; situations where a person is being asked for a sacrifice often carry some sense of moral obligation to comply.

We use debt collection data from a leading online consumer finance company in China that makes uncollateralized installment loans. Borrowers who fail to make their monthly payments on time are contacted on the phone by the company's debt collectors, urging them to repay. The company uses human and AI callers, which allows us to evaluate AI callers' performance relative to humans and to estimate the impact of AI on the company's profits and worker productivity. The AI callers can understand the borrower's speech and generate appropriate voice replies. They provide borrowers with basic information, answer questions, and inform them of the negative

[1] See, for example, Brynjolfsson and Mitchell (2017), Felten et al. (2020), Eloundou et al. (2024), and World Economic Forum (2020, 2023).
[2] https://apps.urban.org/features/debt-interactive-map/?type=overall&variable=totcoll (accessed April 30, 2024).

consequences of defaulting. An important intermediate goal that both AI and human callers try to achieve is to extract a verbal promise to pay from the borrower. Previous research has shown that people find promises psychologically costly to break (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; He et al., 2017; Zhou, 2024).

We identify the relative effectiveness of AI callers using two experiments that occurred in the firm, one natural and one intentional. The natural experiment is created because of the company's rule that nearly all newly delinquent debts with remaining principal no greater than 300 yuan (approximately 42 U.S. dollars) are permanently assigned to AI callers, whereas larger debts are transferred to human callers no later than six days after delinquency begins. Therefore, we can identify the effect of permanent assignment to AI versus assignment to humans using a regression discontinuity design, comparing debts just above the 300-yuan threshold to debts just below.

The intentional experiment is created because each month, the company takes a random 10% of newly delinquent debts with remaining principal greater than 300 yuan and assigns a randomly chosen half to be called by AI through the fifth day of delinquency before being called by humans thereafter (the treatment group) and assigns the other half to always be called by humans (the control group). All debts in this 10% subsample are given to human callers on day 6, so the intentional experiment identifies the effect of a short-lived initial exposure to AI callers versus no exposure to AI callers.

We find in the regression discontinuity sample that when AI callers are permanently assigned to a borrower, they consistently perform worse than human callers over horizons up to one year past due, as measured by the net present value (NPV) of collected repayments scaled by the initial overdue balance. The productivity gap between AI and human callers first widens as days past due increase. It reaches its maximum around one month past due, when the NPV of repayments collected by AI callers is 9 percentage points less than that of human callers. The gap slowly narrows afterward but remains around 5 percentage points even one year past due. In addition, the gap is larger for borrowers with lower credit scores. A very stylized model of debt collectors would characterize their task as merely providing reminders and information to borrowers, and perhaps imposing nuisance costs as well—things that AI can do nearly as well as humans. Indeed, Roll and Moulton (2019) find that automated payment reminders decrease severe delinquencies. But the gap in performance between AI and humans and its heterogeneity by credit quality suggest that this stylized model is importantly incomplete.

The randomized experiment shows that replacing AI callers with human callers after a few days mitigates much of the initial underperformance of AI callers. In this subsample, we continue to find that AI underperforms humans, with the NPV gap monotonically increasing to 12 percentage points by day 5. But the gap quickly narrows once human callers take over the AI cases to 2 percentage points on day 10 and 0.8 percentage points on day 30. Interestingly, the remaining 0.8 percentage point gap barely closes over the next 11 months, indicating that initial contact by AI *permanently* impairs the ability of the company to collect. There may be something uniquely damaging about being contacted by AI that even 355 days of subsequent human contact cannot undo. Repayment reductions resulting from initial contact by a less effective human call (calls on the weekend or by an inexperienced human caller) are mitigated within a few days.

AI also creates alienation beyond the initial domain of contact. One might suspect that if a department chair sends her AI to ask a faculty colleague to serve on an onerous committee, that colleague would be less likely to agree to not only the current request by the AI, but also to future requests from the chair, even if they are made in person. In our setting, for each of the following 12 monthly payments due, borrowers previously contacted by AI are 1 to 2 percentage points more likely to be late than borrowers never contacted by AI.

We explore some potential sources of the initial AI performance gap by examining detailed outcomes of phone conversations in the randomized experiment, restricting to phone calls on the first day of contact. Humans call borrowers nearly one more time per day than AI callers. To remove the impact of additional phone calls, we further restrict our sample to the first call answered by borrowers. After controlling for the call's time of day, we find that AI callers are not more likely to be hung up on within 10 seconds, but they have conversations that are 31 seconds shorter on average and exhibit less variability in length, suggesting that AI callers are less able than humans to provide information, be engaging, and/or handle complicated situations. Moreover, 21 percentage points fewer borrowers promise to repay their debts and 18 percentage points fewer repay on the same day as the call if they talk to AI callers. Conditional on making a promise to repay, borrowers are less likely to keep such a promise when it is made to an AI. Therefore, AI callers appear to be worse than humans at extracting promises that feel binding. This may be because machines are not thought to be owed moral duties (Melo, Marsella, and Gratch, 2016; Petisca et al., 2020). To the extent that promises to AI do have moral force, it is less unpleasant to be commit a moral transgression in the presence of a machine than a human (LaMothe and Bobek, 2020; Kim et al., 2023).

We next consider how AI affects the productivity of human callers. During our sample period, the AI experienced five upgrades, mainly improving its speech recognition and understanding. Each upgrade was rolled out so that two versions of AI callers were used simultaneously in the same month and assigned cases at random. This arrangement allows us to measure improvements in AI productivity and their impact on human callers' performance. We observe that the AI most significantly improved between August and October 2021, increasing NPV collected through day 5 by 3 percentage points. However, using a better AI for the first five days leaves more difficult cases to day 6, so human callers on day 6 collect 3 percentage points less, resulting in similar cumulative collected NPVs. This finding is consistent with a displacement effect of AI on labor.

Finally, we examine whether AI is more cost-effective than humans net of labor costs. We focus on direct labor costs, i.e., workers' salaries, which consist of a fixed component and a variable component. Although the productivity deficit of AI is diminished once labor costs are accounted for, AI remains less cost-effective than human callers except among smaller debts, where AI matches humans. Importantly, this calculation does not consider indirect labor costs, such as recruitment, training, management, pension funds, etc., nor the cost of developing the AI software.

Our paper is related to prior work on negative human reactions to AI. Luo et al. (2019) find that AI callers can be as effective as proficient human callers at persuading borrowers to take up an attractive loan renewal offer, but AI's success rate falls by 80% if it begins the sales conversation by disclosing that it is an AI. Customers rate the AI caller to be less knowledgeable and empathetic when the disclosure occurs up front. Our work is distinct from Luo et al. (2019) in that we study AI's ability to persuade humans to undertake an *unappealing* action and examines the long-term reaction to AI and its spillover to other choices instead of only a single response that comes within one minute. Silva, Khera, and Schwamm (2024) report that editorial board members are unable to identify which persuasive essays in their field of expertise are authored by AI versus humans, but they judge an essay to be lower-quality if they believe it to be AI-generated. Glikson and Woolley (2020) survey the literature on the determinants of human trust in AI.

Our paper is also related to the literature on the impacts of automation on labor. Previous studies find different impacts in different waves of automation.[3] They mostly find complementarity between humans and AI when AI only provides predictions and suggestions and human workers make the final decision (Gao and Jiang, 2021; Luo et al., 2021; Brynjolfsson et al., 2023; Noy and Zhang, 2023). In contrast, the company in our study delegates either all or none of a phone call to AI, since it is hard for AI to assist human callers in real time during conversations. In such a setting, we find imperfect displacement effects; AI callers can replace humans but are less productive by themselves, and they do not make humans more productive when working in tandem.

Additionally, our study contributes to the literature on the performance of AI and machine learning technology (Cao et al., 2024; Erel et al., 2021; Kleinberg et al., 2018; Agrawal et al., 2019, 2023). We focus on non-routine jobs, which were previously believed to be immune to automation (Brynjolfsson and Mitchell, 2017, Felten et al., 2020) and were rarely studied in this context until recently (Gao and Jiang, 2021; Brynjolfsson et al., 2023; Noy and Zhang, 2023).

Finally, our paper contributes to an emerging literature on delinquent debt collection, a function that directly affects many individuals around the world. Drozd and Serrano-Padial (2017) and Fedaseyeu (2020) examine how variation in debt collection effectiveness driven by information technology and regulations affects credit supply. Fedaseyeu and Hunt (2015) model how reputation concerns drive the usage of third-party debt collection. Cheng et al. (2021) study how consumers fare when taken to court by debt collectors. Zhou (2024) find that a machine learning algorithm that chooses which borrowers are called by human collectors increases repayment rates relative to when these choices are made by human collection officers. Laudenbach and Siegel (2023) address the importance of personal communication in collecting loan repayments. They show that phone calls to late borrowers from bank agents are more effective than mail reminders, and bank agents with more likeable voices are especially effective.[4]

The remainder of the paper proceeds as follows. Section 1 provides institutional background about the company, its debt collection process, and its human and AI callers. Section 2 describes our data, and Section 3 specifies our experimental setups. Section 4 estimates the performance gap

[3] In the early AI and information technology revolutions, some researchers find displacement effects for low-skilled workers and increased demands for high-skilled workers (Acemoglu and Restrepo, 2020, 2022). Others find that new automation technologies are labor-augmenting (Michaels et al., 2014; Tan and Netessine, 2020).
[4] Karlan et al. (2015) also document the importance of a personal touch in text messaging in debt collection. They find that only messages with the names of the bank managers who previously serviced the clients can improve repayment.

between AI and human callers. Section 5 explores the potential sources of the performance gap. Section 6 examines the impact of AI on human labor, and Section 7 concludes.

1 Institutional Background

1.1 The company and its lending business

The company is a leading online consumer finance service provider in China. At the end of 2022, the company had around 10 million active users with nearly 1 trillion yuan (14 billion USD) of outstanding loan balances. The company's main business is to originate loans to online consumers. The company targets young consumers with a short credit history but large income and consumption growth potential.[5] It operates its own online shopping platform and collaborates with third-party online retailers to offer loans at the point of sale.

The 10th percentile loan size is only 8 yuan (1 USD) and the 90th percentile is around 5,500 yuan (770 USD). The company provides two types of loans. The first is an installment loan, which the consumer repays in equal monthly installments over the next six months to three years. The second is a credit-card-like product. Consumers may apply for a credit line, which is around 7,500 yuan (1,050 USD) on average, and pay for their online order with it. "Credit card" loans have a default maturity of one month, but the borrower may choose a longer maturity up to a limit determined by the company's risk assessment of the borrower, in which case monthly payments of equal size are due during the loan's life. "Credit cards" are typically used for small payments, while installment loans are preferred for expensive purchases and durable goods, similar to the setting studied by Aydin (2022).[6] Since the company's customers are typically riskier than the population average, the interest rates are mostly 24% per annum, which is the upper limit allowed by Chinese regulators.

Each borrower is assigned a monthly repayment due date (e.g., the 8th day of each month), which may be changed by the borrower with the company's approval. Changing the due date frequently is not allowed. Borrowers who fail to pay enter the debt collection process. During

---

[5] Seventy percent of the company's customers are less than 30 years old, 65% are members of the urban working population, and 13% of them have a bachelor's degree or more. These percentages are much higher than the population-wide averages.

[6] Extremely small loans are mainly generated by "credit card" purchases. The company views each purchase, no matter how small it may be, as a standalone loan. If we aggregate all newly originated loans from both products to the borrower-month level, the monthly borrowing amount of a customer ranges from 500 yuan (70 USD) at the 10th percentile to 14,000 yuan (2,000 USD) at the 90th percentile.

delinquency, extra interest and late fees accrue on the overdue amount. Borrowers need to repay the overdue amount, accrued interest, and late fees to fully resolve their delinquency.[7] Borrowers who remain delinquent for ninety days are considered to have defaulted and are reported to third-party credit report aggregators. Defaulted borrowers cannot borrow from the company again and may have trouble borrowing from other consumer finance companies. Defaulting may also impair the borrower's ability to make purchases even without borrowing.[8]

1.2 Human debt collection process

The company treats the first day past due as a grace period. It generally does not call delinquent borrowers on that day and just sends them reminders through text messages and phone app notifications. If the debt remains unpaid on the second day past due, the company starts calling.

The company uses different strategies depending on whether the debt is 2-10, 11-25, 26-59, 60-84, or 85+ days past due. During days 2-10, there are three blocks of time when borrowers who have not repaid (excluding those who have verbally promised to pay during a prior call and have not passed their promised deadline) are called automatically by the system: 9 to 9:30 A.M., 3 to 3:30 P.M., and 7:15 to 7:35 P.M. Human callers do not make any calls themselves, but wait for the automatic calls to be answered. When a call is answered, it is sent to a caller who is randomly selected from those available at that time.

Outside the three automatic call time blocks, human callers can choose, based on debt and borrower characteristics shown on their computer screen, which borrowers to call between 9:30 A.M. and 8 P.M. from among those whom they have been randomly assigned (at the beginning of the day) for the day. Callers' performance pay is based on how much they collect from their assigned borrowers. No borrower is simultaneously assigned to more than one caller. According to the company's internal research, the characteristic used most by productive callers to filter cases

[7] Extra interest accrues on the overdue amount at the same rate as the original loan. Late fees accrue on the overdue amount at 0.1% a day (or 36.5% annually). The company may forgive these charges in part or in full for borrowers over 10 days past due as an incentive to repay. Total undiscounted late charges paid by the mean (median) delinquent borrower are 0.7% (0.2%) of the initial overdue amount. A few severely delinquent borrowers paid 10% to 30% of their initial overdue amount in late charges.

[8] For example, defaulted borrowers may not be able to use rideshares or book hotels without deposits, since some large companies use credit records for screening. If the lender can prove that the borrower is not repaying despite having enough money, it can sue the borrower. If the lawsuit is supported by the court but the borrower still refuses to repay, the borrower will be added to a blacklist of "dishonest judgment debtors" assembled by the Supreme People's Court of the People's Republic of China and prohibited from expensive consumption such as traveling by plane and purchasing real estate and luxury cars.

is the most recent time that the borrower logged into the company's app. The research also suggests that case selection skills play only a minor role in explaining human callers' performance. To prevent the company's phone numbers from being blacklisted by borrowers, the company uses multiple phone numbers, and the caller can choose which one will be displayed on borrowers' caller ID.

During a phone call, callers usually provide information about the loans, inform the borrower of the potential negative consequences of delinquency, and try to persuade them to repay the debts as soon as possible. Callers are provided some conversation templates but are not asked to follow them strictly. Callers may provide suggestions to borrowers, such as encouraging them to ask family members for help. These suggestions are typically short and generic in days 2-10. In later stages, the conversations are more personalized and specific.

Callers have an incentive to ask borrowers to make payment promises. If in a conversation, the borrower clearly and explicitly states that they will repay the amount due no later than the end of the next day, the caller will label it as a "promise to pay," and the case will be kept by the same caller for one more day. Cases for which neither a promise nor a full repayment has been made get reassigned to a different caller each day. If a repayment promise is made to a caller during an automatic call block, ownership of the borrower's case is transferred to this caller until the end of the following day. Repayment is credited to the caller who received the promise as long as it comes within the promised time. The company uses AI to identify false "promise to pay" labels, which are double-checked by human examiners. Each false label results in a 300-yuan salary deduction, which is large relative to the average caller monthly salary of around 5,000 yuan.

At later stages, the company uses different debt collection strategies. A caller will by default handle a debt in days 11-25 for one week before it is assigned to another person. For debts in days 26-84, this interval is typically two weeks. A caller who extracts a payment promise can hold on to the case for 1-3 days longer than normal, but not across the boundary between days 25 and 26 or the boundary between days 84 and 85. Automatic calls are made four times a day during days 11-25 but cease thereafter. Beyond 84 days, we have less information about strategies, since nearly all cases are handled by third-party debt collection agencies. Third-party agencies are rarely used before day 26, except for a few months in which in-house callers were overloaded and around 10% of overdue debts were delegated to third-party agencies between days 20 and 25. Around 60% of overdue debts are outsourced to third-party agencies during days 26-59, and around 80% are outsourced during days 60-84. The company sends almost all 85+ day late cases to third-party agencies, keeping only small cases to be collected by AI callers (see Section 3) and some very large ones for further actions like lawsuits. Borrowers are typically contacted less intensely as the repayment gets further past due (see Online Appendix Figure B1).

1.3 AI caller

To cope with the high volume of cases and to reduce labor costs, the company introduced AI callers in 2018. The rules determining whether a borrower is assigned to AI will be explained in Section 3. The AI callers understand borrowers' speech and generate appropriate answers, speaking with a synthetic voice which is easily recognized as not human. The AI can provide basic information about the overdue loans, address potential negative consequences of delinquency, and respond to simple questions.

Table 1 illustrates the conversation process and some sample scripts that the AI caller typically uses. These scripts do not change as days past due increase. The conversation is divided into four stages by design. In the first stage, the AI greets the borrower and confirms their name. If the AI dialed the wrong number, it tries to confirm the relationship between the call recipient and the borrower and asks the recipient to convey a message to the borrower if the two know each other; otherwise, the AI apologizes and hangs up.

If the call recipient is the borrower, the AI continues to the second stage to inform the borrower about the overdue debt. The information provided by the AI includes the overdue amount and the number of days past due. The borrower is usually asked to repay within two hours.

The AI then waits for the borrower's responses. The software classifies possible responses into seven broad categories. In the first case, the borrower agrees to repay within two hours. The AI then tells them that their promise has been recorded and may ask them to confirm their promise or to repay within the promised time. In the second case, where the borrower asks for an extension, the AI will compromise and ask the borrower to repay by the end of the day (but never later than that). In the third case, the borrower is unable to repay the debt and may explain why. The AI can understand these explanations and reply accordingly. For example, for liquidity problems, the AI may ask the borrower to borrow money from their family or friends. To borrowers who claim to be too busy to pay, the AI may say that it understands that they are busy but urges them to spend just a few minutes to repay the debt online. In the first three cases, the AI also emphasizes potential negative consequences if the borrower fails to repay: more phone calls, worsening credit records, escalating debt collection actions, late fees, and difficulties in future borrowing and consumption. The AI may also mention the possibility of informing the borrower's "emergency contacts," who are typically their parents and colleagues, imposing social pressure.[9] In the fourth, fifth, and sixth cases, the borrower claims that they do not have any debt with the company, have already repaid, or have set up auto-payment. The AI then asks the borrower to recall their borrowing history, to double-check their accounts, or to make sure they have enough money in their account to make the auto-payment. In the final case, the AI answers questions about basic information about the debts, such as late fees.

Finally, when the borrower has no more questions about their loans, the AI may reiterate the negative impacts of delinquency and will ask the borrower to contact customer service for further information. Similar closing words are also used to end the conversation when the AI cannot recognize the borrower's responses due to loud noises, strong accents, etc. (0.4% of calls), when there is a long silence (22% of calls), or when the borrower's responses cannot be classified into one of the pre-specified cases (for example, the borrower yells at the caller or complains about the annoying phone calls).

Online Appendix C contains some sample AI call transcripts.

2 Data Description

Our data provide us with comprehensive information about the debt collection process in the company between April 2021 and December 2023. To ensure that we can track each delinquent debt for at least one year, we restrict our analysis to cases entering collection before December 2022, which gives us more than 22 million cases. Consistent with the company's practice, multiple delinquent debts of an individual borrower are merged into one entry during collection.

We have loan and borrower characteristics for all delinquent debts, including loan size at delinquency, borrower internal credit score, age, gender, and education level. The company uses two different measures of loan size: the overdue payment amount and the remaining principal. The internal credit score is based on the probability of default estimated by the company. The company divides all delinquent borrowers into deciles and assigns them an integer score from 1 to 10, where

[9] Each borrower is asked to list two "emergency contacts" as their "guarantors" at the time of registration. These "emergency contacts" do not have any legal obligation to repay the loans if the borrower defaults, although some emergency contacts may be willing to repay the borrower's debt. Calling emergency contacts is often viewed as an escalation of the debt collection actions and is tightly restricted by the regulator.

1 is the highest decile of default probability. This score is updated daily, incorporating the phone call outcomes of the previous day and the daily loan sizes. Education levels are self-reported, although the company can verify some of them if borrowers have uploaded their degree certificates and transcripts when registering their accounts.

We also have daily records of debt collection status and repayment actions. We know the number of days overdue, the caller handling the loan, and when and how much the borrower repays each day. We additionally have information about every phone call made to borrowers, including the time of each phone call, the caller, how long the borrower's phone rings, whether the phone call is answered by the borrower, whether the borrower has been labeled as having promised to repay during the conversation, and the duration of the phone call.

Finally, we have data about demographics of all callers, and monthly performance and compensation for in-house callers. Callers' demographic information includes their age, gender, city of birth, and whether they are in-house or with a third-party collection agency. For in-house callers, we have their job titles and their tenure (in months) with the company. Performance measures for in-house callers include the total amount of money collected, monthly target collection amount, performance ranking, and the ratio of the actual amount collected by the caller to her target. We know the salary amount each in-house caller received, as well as the portion that is performance-based and the amount that is deducted due to penalties.

Table 2 Panel A reports summary statistics for the loan and borrower characteristics in the full sample of delinquencies. The characteristics are measured on day 2 past due, the first day when cases enter the debt collection process. The average delinquent debt has an overdue amount of 1,128 yuan (160 USD) and a remaining principal of 6,474 yuan (910 USD), which are larger than the corresponding moments for the population of outstanding loans, as larger loans are more likely to default. The medians are smaller than the means: the median overdue amount is 654 yuan (92 USD), and the median remaining principal is 4,248 yuan (600 USD). The average internal credit score is around 5. Among delinquent borrowers, 70% are males, 13% have a bachelor's degree or more, and the average age is 27 years.

Case sizes are heavily right-skewed: the maximum remaining principal is 1 million yuan (about 140,000 USD). Extremely large debts are typically nonstandard contracts with specific customers for special purposes. They are treated separately by the company, so we want to exclude them from our analysis. Since separately treated cases are not labeled in our data, we exclude cases with remaining principal above the 99th percentile. The left tail does not require trimming since extremely small cases are excluded in our experimental design, as discussed in the next section.

3 Experimental Setups

To identify the productivity difference between AI and human callers, we utilize the company's rules for assigning cases between AI and human callers. Figure 1 illustrates the assignment procedure. First-time delinquent borrowers are always assigned to human callers. Starting with the second delinquency, borrowers can be assigned to either AI or human callers.

The company initially allocates all cases with overdue amounts no greater than 20 yuan or remaining principal no greater than 300 yuan to AI callers.[10] In rare situations, which we discuss later in this section, these small cases are assigned to human callers after day 25.

Larger cases are either unconditionally or conditionally randomly assigned to AI or human callers. The company randomly selects 10% of larger second-delinquency cases every month for testing and monitoring purposes.[11] In this subsample, a random half of cases are assigned to human callers on day 2, while the other half are assigned to AI callers on days 2 to 5 before being reallocated to human callers on day 6 onwards. Once a given delinquency is handled by a human caller, it typically will not be given back to an AI caller.

For subsequent delinquencies, the borrower's assignment to be initially called by a human or an AI remains the same as it was for his second delinquency unless the subsequent delinquency is small (overdue amount $\leq$ 20 yuan or remaining principal $\leq$ 300 yuan), in which case it always gets assigned to AI. Therefore, only the assignment in the second delinquency can be viewed as orthogonal to potential outcomes within our sample; the type of person who reappears in our data as delinquent a third time after always being called by a human might be different on average from the type who becomes delinquent a third time after being called by AI. Thus, for larger cases, our analyses focus only on borrowers in their second delinquency, which are about 11% of the full sample. We call this subsample of second delinquencies the "completely randomized subsample."

---

[10] If the borrower with a debt size above this threshold makes a partial repayment that brings the debt below the threshold, the case is switched to being handled by AI. Of cases in the (300, 305] yuan day-2 remaining principal interval, 1.67% were ever switched from being initially handled by humans to being handled by AI because of a partial repayment between days 2 and 360.

[11] Larger cases that enter the debt collection process on the last few days of each calendar month are always assigned to human callers because there are fewer cases initiated at the end of each calendar month. These cases are excluded from our analyses. Borrowers who are delinquent again retain their prior assignment to the unconditionally or conditionally randomized group.

The remaining 90% of larger second-delinquency cases are assigned between AI and human callers randomly conditional on case characteristics; that is, the probability of a case being assigned to AI varies by its characteristics, as does the timing of when humans take over from AI (anywhere from day 2 to day 6). AI treatment effect estimates within this conditionally randomized subsample are similar to those in the completely randomized subsample, so we do not report them.

Whereas the completely randomized subsample allows us to identify the effect of replacing human callers with AI callers from days 2 to 5, the discontinuity in the company's assignment rule for small cases creates an opportunity to use a regression discontinuity (RD) design to identify the local treatment effect of replacing humans with AI for a much longer time. The 20-yuan overdue amount threshold is extremely small—almost at the 1st percentile of the full-sample distribution. The 300-yuan remaining principal threshold is somewhat less extreme—at around the 5th percentile of the full-sample distribution. Therefore, in the RD analysis, we exclude cases with less than 20 yuan of overdue payments and apply the standard RD methodology with one running variable, the remaining principal, to the sample of second-time and later delinquencies.

Figure 2 Panel (a) plots the fraction of second-time and later delinquencies assigned to AI as a function of the remaining principal on day 2. Consistent with the stated assignment rules, cases below 300 yuan of remaining principal are all assigned to AI callers, while only about 80% of cases above the cut-off are assigned to AI. The discontinuity in the AI fraction is sharp.

Figure 2 Panel (b) shows the fraction of cases assigned to AI callers on either side of the threshold from days 1 to 25. The fractions for "Under 300" are calculated based on cases in the (295, 300] yuan interval, while the fractions for "Above 300" are calculated based on cases in the (300, 305] yuan interval. Small cases are all handled by AI callers in the first 25 days. In contrast, on days 2-3, only 80% of the larger cases are assigned to AI callers. The fraction falls to around 60% on days 4-5. From day 6 onwards, all larger cases are handled by human callers. Panel (c) extends the horizon to day 360. Cases above 300 yuan remain under human treatment for the entire extended period. For cases below 300 yuan, a small fraction of them are assigned to human callers after day 25, mainly due to the introduction of third-party collection agencies. When the company delegates to a third-party agency, it randomly selects some cases, maybe conditional on some loan characteristics, and assigns them to the agency. The assignment of some small cases to humans biases against finding significant collection differences across the remaining principal threshold.

4 AI versus Human Caller Performance

4.1 Measure of debt collection productivity: Net present value of collected cash flows

We use the net present value (NPV) of cash flows collected from day 2 past due onwards as the measure of caller productivity. For each case, we calculate how much money is paid towards the initial delinquent payment on each day, including late fees. We then discount these cash flows to day 2 using a 24% per annum (24%/365 = 0.066% per day) discount rate, which is close to the average APR of the loans originated by the company. It is also the maximum legal APR allowed by Chinese regulators. It can be viewed as the opportunity cost of uncollected money, which could have been lent to other borrowers and generated interest at a 24% APR if it were collected on time.[12] Finally, the NPV is scaled by the amount overdue on day 2. The scaled NPV can grow above 1, since borrowers may pay late fees (see footnote 7).

4.2 Small cases subsample: Regression discontinuity design

In this subsection, we compare the productivity of AI callers to human callers by utilizing the discontinuity in the company's AI deployment strategy at the 300-yuan cutoff in remaining principal.

Table 2 Panel B reports summary statistics for loan and borrower characteristics in our subsample for the RD design: cases in their second or later delinquency with remaining principal between 100 and 500 yuan, which gives us over 1 million cases. Although loan sizes are much smaller than in the full sample, as expected, the gender composition, average age, and the fraction of borrowers with a bachelor's degree or more are all close to those in the full sample. The average internal credit score is somewhat lower in the RD sample than in the full sample but is still very close to 5.

To assess the validity of our RD design, Table 3 Panel A reports continuity tests on five predetermined loan and borrower characteristics around the 300-yuan remaining principal threshold.[13] The regression-fitted average loan characteristics on the two sides of the cutoff

---

[12] Setting the discount rate to zero has little impact on the results, since most payments are collected in the early days of delinquency.

[13] Local linear regressions with uniform kernels over the coverage error rate (CER)-optimal bandwidths are used in the estimation, and robust bias-corrected RD z-statistics estimated by local quadratic regressions are reported. As suggested by Cattaneo et al. (2019), the CER-optimal bandwidth is used for testing the null hypothesis of continuity because we are interested in inference (the confidence interval) instead of point estimates. The authors also show that the degree of local polynomials for inference should be one order higher than the degree of local polynomials for point estimates for a bias-corrected statistic that is robust to bandwidth selection.

(columns 2 and 3) are quite similar, and their differences are small (column 4) and not statistically distinguishable from zero (columns 5-7). Figure 3 Panel (a) shows these tests graphically. In Online Appendix A, we check if there is manipulation around the cutoff by examining the density of observations. Although the remaining principal amount has some tendency to cluster at 300 yuan (and also at 200 and 400 yuan), the density functions can be considered continuous at the threshold.

Having established that our empirical design is likely to be valid, we measure the difference in the average collected NPV between the two sides of the 300-yuan cutoff, which gives the treatment effect of AI callers on debt collection productivity. Table 3 Panel B presents estimates of the NPV differences at various horizons.[14]

The mean NPV estimated from the left (AI) minus the mean NPV estimated from the right (Human) of the cutoff is negative and significant at the 1% level, regardless of the evaluation horizon. These gaps are also economically significant. On day 2, the NPV gap of 0.04 is a 19% productivity loss relative to the human mean NPV of 0.21. The gap grows to 0.09 by day 30 before starting to shrink because human callers do not collect much more beyond day 30, whereas AI continues to make some significant collection progress. Nonetheless, even after 360 days, AI's productivity loss relative to humans remains large: 5%. Figure 3 Panel (b) presents the RD plots of collected NPVs through various horizons after the initial due date. Figure 4 shows the productivity gap over time graphically. Re-estimating the treatment effects while controlling for the five loan and borrower characteristics in Table 3 Panel A has a negligible impact on the estimates (Table 3 Panel B column 8).

One potential explanation for low AI productivity early in delinquency is that humans call borrowers two to four times as often as AI does during the first 25 days (Online Appendix Figure B2). However, we will later see in the completely randomized subsample that AI underperformance early in delinquency is large even when call frequencies are much more similar between humans and AI. Later in delinquency, human caller effort drops considerably to a level below AI's; AI continues to call each still-delinquent borrower almost twice a day through day 360. This persistence of effort could contribute to AI making up ground on human callers after day 30.

Next, we examine how AI's performance deficit varies with borrowers' credit quality in Figure 5, using the same specification as in Figure 4. Low, medium, and high groups refer to internal

---

[14] Since we are now interested in point estimates of the productivity gap, the mean squared error (MSE)-optimal bandwidths are used in these regressions.

credit scores of 1-3, 4-7, and 8-10, respectively, measured on day 2. AI initially underperforms more with high-score borrowers. However, the gap between AI and humans shrinks quickly for the high group, approaching –2% in the long run. In contrast, the performance gaps for low-score borrowers keep expanding until around 30 days. The magnitudes of the long-run productivity gaps are monotonically decreasing in credit scores. High-score borrowers may mainly need reminders, which AI callers can provide adequately, while low-score borrowers likely find repaying more costly and may therefore require greater persuasion, which AI callers are less capable of performing.

4.3 Completely randomized subsample

The previous subsection shows that AI callers alone are less productive than human callers alone. The company usually has AI callers supplement human caller efforts in the early stages of delinquency. Specifically, some cases are assigned to AI callers at first. If the delinquency has not been cured soon, human callers take over.

To identify the performance of the "AI + Human" strategy, we utilize the completely randomized subsample. In this subsample, the company randomly selects half of the cases and assigns them to human callers from day 2 onwards (the control group), while the remainder are assigned to AI callers on day 2 and reallocated to human callers from day 6 onwards (the treatment group). Table 2 Panel C shows summary statistics on the completely randomized subsample. Since small cases with remaining principal no greater than 300 yuan are excluded from this subsample, the overdue amount and remaining principal here are on average larger than in the full sample. Other borrower characteristics are similar to the full sample.

As a first step, we validate that the treatment and control groups are comparable to each other. We regress predetermined loan and borrower characteristics onto a treatment group indicator and dummies for the calendar month of the second day of delinquency. Table 4 Panel A shows that the coefficient on the treatment group indicator is insignificant when the dependent variable is overdue amount, remaining principal, internal credit score, gender, age, and education level. This shows that the two groups are statistically indistinguishable from each other ex ante.

We then estimate the productivity gap between "AI + Human" and the always-human control using the same regression specification. Results are shown in Table 4 Panel B. Columns 2 and 3 show the average cumulative NPV collected of the treated (AI + Human) and control (Human) groups, respectively. Column 4 reports the difference between the two groups, and the next column reports the *t*-statistic of this difference.

For all evaluation horizons, the "AI + Human" treatment group significantly underperforms the always-human control group. The gap is 0.09 on day 2, the first day of contact, which corresponds to a 33% productivity loss relative to the always-human control, and expands to 0.12 on day 5, a 24% productivity loss. Once human callers take over after day 5, the performance of the two groups converges quickly, so that the NPV difference is only 0.02 on day 10 and 0.008 on day 30. Nevertheless, the "AI + Human" group never repays as much as the control group; even after a year, the gap is 0.007. On the one hand, this is only about a 1% relative productivity loss. On the other hand, it is remarkable that only four days of exposure to AI callers permanently impairs the company's ability to collect. The last column re-estimates the differences with loan observations weighted by their initial overdue payment amount on day 2. These value-weighted differences, which represent the total monetary losses from AI treatment, are larger in the early days but similar in the long run. Figure 6a presents these results in graphical form.

Figure 6b and Figure 6c plot the NPV gap over time by internal credit score and loan size, respectively, measured on day 2. The low-score cases suffer the least productivity loss from AI initially, which is similar to what we saw in the RD analysis. We also learned from the RD design that the NPV gap of low-score cases would keep growing and exceed the gaps of the other two groups if AI callers continued working on them. In the completely randomized subsample, however, human callers intervene on day 6, halting the damage. Therefore, low-score cases also have the least performance damage over longer horizons. High-score cases experience the largest permanent damage from initial AI contact. On the loan size dimension, cases with larger overdue amounts generally have larger performance gaps initially, consistent with our expectation that they require more persuasion that AI is less able to handle. In the long run, larger loans are more damaged by initial AI contact than smaller loans, but mid-sized loans with overdue amounts between 800 and 1500 yuan are the most damaged.

5 Understanding AI's underperformance

5.1 Calls to "emergency contacts"

Why does AI underperform humans in collecting payments? One potential reason is that human callers are more likely to call the borrower's "emergency contacts," such as their parents and colleagues, creating an additional source of social pressure to repay. Online Appendix Table B1 indicates that in the regression discontinuity subsample, the fraction of calls that go to emergency contacts is actually 1 percentage point higher for AI-only cases just below the 300-yuan remaining principal threshold than for the 20% AI/80% human cases just above the threshold in days 2-4. Recall that collections from AI-only cases are nonetheless significantly lower than from the 20% AI/80% human cases through day 5.

After day 5, emergency contacts of larger cases start persistently receiving more calls. On day 6 past due (when humans start handling all larger debts), calls to emergency contacts are 4 percentage points higher just above the 300-yuan threshold than just below, 16 percentage points higher on day 10, 14 percentage points higher on day 60, 7 percentage points higher on day 90, and 1 percentage point higher on day 360.

We can gain a sense of how much of the long-run human advantage in collections is coming from a higher propensity to call emergency contacts by examining the completely randomized subsample. Figure 7 shows that prior to April 2022, both AI and human callers in this subsample call emergency contacts less than 3% of the time on day 2, but by day 4, humans call emergency contacts 19 percentage points more often than AI. Starting in April 2022, the company changed its policy so that neither humans nor AI made any calls to emergency contacts before day 5.

In untabulated results, we find that AI collected 9.4 percentage points less NPV than humans through day 4 (t-stat = -23.5) from April 2022 onwards. Recall too that in Table 4, AI collects 8.8 percentage points less NPV than humans on day 2, when calls to emergency contacts almost never happen throughout our sample period. We conclude that AI is less productive even when neither humans nor AI calls emergency contacts.

However, AI's performance deficit through day 4 prior to April 2022—13.3 percentage points (t-stat = 44.4)—is larger than it is from April 2022 onwards, suggesting that calling emergency contacts is a potent strategy that contributes to some of the collection gap between AI and humans over the entire sample period.

5.2 Underperformance on day 2

To gain insight into why AI underperforms humans even when both refrain from calling emergency contacts, we examine other outcomes of phone calls made on day 2, when calling emergency contacts is always rare—the duration of the call, the fraction of calls that result in a promise to repay, and the fraction of promisers who make payments shortly after the call.

Table 5 Panel A reports the average outcomes of all phone calls made by AI and human callers on day 2 within the completely randomized subsample. Human callers make 1.38 (29%) more phone calls per day to each borrower than AI callers and thus are answered 0.35 more times per day. Of phone calls from human callers, 16.3% are answered, which is 2.6 percentage points higher than for phone calls from AI callers.

To analyze differences in the ability of AI versus human callers that are separate from the frequency with which they make calls, we next restrict our sample to the first call answered by each borrower. The results are reported in Table 5 Panel B. The time of first-answered calls is on average a little earlier for human callers than for AI callers: 11:31 AM versus 11:47 AM. The reason is that, as mentioned in Section 1.2, there is a half-hour automatic call period from 9 A.M. to 9:30 A.M. when all cases without a still-active payment promise that are assigned to human callers are called once. In contrast, calls from AI callers are distributed more evenly across the day. To control for this disparity, we estimate timing-adjusted results that control for one-hour-interval time-of-call fixed effects.[15]

There is a significant 1.3 second difference between the two types of callers in how long the phone rings before it is answered, but this disparity disappears after controlling for hour-of-call fixed effects. On the other hand, the duration of phone calls significantly differs whether or not hour-of-call fixed effects are controlled for. The mean duration of an AI call is only 28 seconds, which is 19 seconds less than for calls by human callers. The gap widens to 31 seconds after the timing adjustment. This finding suggests that AI callers are less able than humans at providing information, being engaging, and/or handling complex situations, leading to short conversations.

Online Appendix Figure B3 shows the histograms of phone call durations for the two types of callers separately after reweighting the calls so that the hour-of-call distributions of the two caller types match. One potential interpretation of the differences in average call times described above is that borrowers hang up quickly upon realizing that an AI is calling. However, the figure

[15] The difference in average call time of day is small, but strictly speaking, a borrower who first answers a human call at 11:00 A.M. has missed more calls on average than a borrower who first answers an AI call at 11:00 A.M., so controlling for time of day may not control for all unobservable borrower characteristics. In Online Appendix Table B2 Panel A and Online Appendix Table B3, we restrict the sample to the first call attempted between 9 and 9:30 A.M. The results are similar: AI callers receive fewer promises to repay and these promises are less likely to be kept by the end of the day. Within the first 15 minutes, more borrowers in this subsample repay their debts when talking to an AI than to a human. Appendix Table B2 Panel B splits the sample by the overdue payment amount at 1,000 yuan (around the median size) and finds that this 15-minute pattern is mostly driven by small borrowers, with whom human callers have especially short conversations on average.

shows that AI calls are almost equally likely to terminate within the first 10 seconds as human calls, which suggests that immediate hangups do not explain the difference between AI and human caller performance. AI phone call lengths are concentrated around 30 seconds, while the duration of human calls has greater variation—potentially a proxy for flexibility in response to complexity.

We additionally find that AI callers are 21 percentage points less likely to receive a promise to repay.[16] This result must be interpreted carefully because AI callers classify a call as having resulted in a promise if they believe there was a promise to pay on the *same* day as the call, whereas human callers classify a call as having resulted in a promise if there was a promise to pay no later than the end of the *following* day.[17] Nonetheless, we see in the first two columns of Table 6 that within one hour of the call—well before the end of the following day—borrowers who made a promise to a human are 3 percentage points more likely to have repaid than borrowers who made a promise to an AI. By the end of the same day the call was made, this difference is 20 percentage points. Thus, additional promises received by human callers do not come entirely from borrowers who plan on paying the following day, inconsistent with the hypothesis that humans are able to extract promises from more borrowers simply because the deadline for a payment promise to a human is later than the deadline for a promise to an AI. Relatedly, one might have expected that if AI callers are bad at extracting promises, they would disproportionately receive promises from borrowers who are likely to pay promptly anyway—the low-hanging fruit. But borrowers are *less* likely to pay after a promise made to an AI than after a promise made to a human.

It is also interesting that in the absence of a promise, the difference in collection probability between AI and human callers is muted: 0.1 percentage points higher for AI within one hour of the call and 6 percentage points higher for humans by the end of the day of the call. Borrowers who make a payment promise to an AI are 13 percentage points more likely to pay by the end of the day than borrowers who speak to an AI but do not make such a promise, which is only half of the 25-percentage-point promise versus no promise difference for borrowers who speak to a human. Keeping in mind the caveat that borrowers endogenously choose whether to make a promise in a

[16] In this calculation, calls to emergency contacts are never classified as resulting in a promise to repay.

[17] To check the accuracy of the AI's classification, a native Chinese-speaking research assistant read 200 AI call transcripts, blinded to the AI's classification, and classified whether the borrower made a promise to pay by the end of the day. Out of 100 transcripts where the AI judged a promise to have been made, the research assistant labeled 84 as containing a promise (80 to pay on the same day, 4 to pay by the next day). Out of the other 100 transcripts where the AI judged no promise to have been made, the research assistant labeled 95 as containing no promise. So the AI probably overestimates the prevalence of promises by $0.16 \times 44.1 - 0.05 \times (100 - 44.1) = 4$ percentage points.

manner that may be correlated (differently depending on the caller type) with their propensity to pay without having made a promise, these numbers suggest that a promise to an AI creates less of a sense of obligation than a promise to a human. Table 5 Panel C shows that integrating across both borrowers who do and do not make a promise, an answered call from an AI is 18 percentage points less likely to result in a same-day payment than an answered call from a human.

How much does the difference in phone call frequency between AI and humans explain the performance gap between them on day 2? We provide suggestive evidence in Table 7. Because call frequency to a given borrower may depend on the borrower's behavior, we avoid reverse causality confounds by not directly examining how borrowers' repayment relates to the frequency with which they are called. Instead, we sort human callers specializing in the first five days past due into terciles based on four different measures of their call frequency to all their cases in the *previous* month. We find that borrowers this month are called more frequently on day 2 if the caller they are assigned at the beginning of day 2 was a more frequent caller in the prior month; the differences between the highest and lowest terciles are statistically significant for all four frequency measures. Borrowers are randomly assigned to callers, so this relationship is due to caller fixed effects rather than borrower characteristics. We then identify the effect of call frequency on borrower repayment by seeing how repayment this month responds to being randomly assigned a caller who called others more versus less frequently in the prior month.[18]

We find that the number of phone calls per borrower-day, the probability of calling again on the same day that a promise is made, and the average minutes until the follow-up call after a promise (conditional on following up on the same day) do not have statistically robust monotonic relationships with payments collected on day 2. Only the average time interval between consecutive calls has significant predictive power for collections on day 2. Specifically, being initially assigned to a human caller in the lowest tercile results in 11 minutes less on average between calls received than being initially assigned to a caller in the highest tercile, which is associated with 1.4 percentage points more collected NPV on day 2. In unreported results, a binned scatter plot of day 2 scaled repayment against the average time between calls indicates that the relationship is close to linear. AI callers take 13 minutes longer between calls than human callers

[18] This identification will be confounded if a caller's call frequency strategy is correlated with unobserved features of the caller that affect borrower repayment. Also, recall that during the three automatic call blocks, borrowers are called by humans that were not assigned to them at the beginning of the day.

in the median tercile, so the time gap between calls only explains 1.7 percentage points of the gap between AI and the median tercile's performance. By comparison, the average productivity gap on day 2 between AI and all human callers is 8.8 percentage points.[19]

In summary, the evidence suggests that while some of the performance deficit of AI relative to human callers can be attributed to AI's lower call frequency, AI also has a lesser ability to extract promises to repay and create pressure to keep those promises. AI additionally appears to be less able to communicate with borrowers, engage their attention, and/or handle complex situations.

5.3 Permanent collection impairment

In Section 4.3, we documented that collection is permanently impaired for borrowers who are first contacted by AI callers instead of human callers. Although it is possible that borrowers inferred from the AI contact that the company was less serious about collecting the debt, it is hard to believe that 355 days of subsequent contact from human callers would not eliminate differences in this belief relative to borrowers who were always contacted by humans.

Table 8 shows that borrowers contacted by AI experience a more general alienation from the company. Not only are they less likely to ever repay the initial delinquent payment due, but they are more likely to miss each of their next 12 monthly payments. Conditional on being delinquent for a second time (and thus entering the completely randomized subsample) and having at least one more monthly payment due before their loan matures, 60% of borrowers always contacted by humans miss that next monthly payment. This probability of delinquency is 2.3 percentage points higher for borrowers first contacted by AI. Among borrowers who have at least 12 more monthly payments due, borrowers first contacted by AI are 1.3 percentage points more likely to miss the twelfth payment due than borrowers always contacted by humans. Thus, whatever damage AI is doing is not narrowly isolated to the payment it initially tried to collect.

To assess whether there is something uniquely damaging about AI, or whether *any* less-productive collection method used in the early days of delinquency also causes permanent

[19] If we take seriously the point estimates of all monotonic relationships between caller behavior and collected money, whether or not the relationship is statistically significant, and extrapolate the relationship estimated among human callers to AI, we would attribute an additional 1.4 percentage points of the AI productivity shortfall to follow-up call behavior towards borrowers who made a promise to repay. Forty-four percent of borrowers make a promise to AI, and within that subgroup, AI's longer time to follow-up conditional on following up accounts for $-0.006/18.77 \times (214.39 - 94.61) = -3.83\%$ less NPV collected relative to the middle tercile of human callers. Weighting by the fraction of borrowers affected, we get $0.44 \times 0.829 \times -0.0383 = -1.40\%$.

impairment, we examine the long-run effect of lower *human* caller productivity in the initial stage of debt collection.

First, we exploit the fact that collection calls that occur on weekends are less effective. Since a debt's due date normally occurs on the same day of each month (e.g. the 11th) and cannot be changed frequently, whether a borrower is first contacted about a late payment on a weekend should be uncorrelated with borrower and debt characteristics. We confirm this orthogonality by regressing overdue payment amount (or its log), remaining principal (or its log), internal credit score, gender, age, or attainment of a bachelor's degree measured on day 2 after the due date on an indicator for if that day is a Saturday or Sunday. We also control for week fixed effects so that only variation within the same week is used for identification. The sample is the completely randomized subsample. In untabulated results, we find that no t-statistic on the weekend dummy has a magnitude greater than 1.29.

In Table 9, we see that borrowers who are first contacted on weekends repay 1.5 percentage points less in normalized NPV on the day of contact than borrowers first contacted on weekdays. However, the gap's significance lasts for only three days. Thus, the productivity loss from first contacting a borrower on a weekend is transient, unlike initial contact from an AI caller.

We next explore variation in caller working experience, measured by the number of months since the caller joined the company. The company assigns debts randomly among human callers every day, so some debts are assigned to more experienced callers at the beginning of day 2. In line with the company's operating and managing practices, we define senior callers as callers who joined the company more than four months ago. We regress variables of interest onto a senior-caller indicator with month fixed effects using debts in the completely randomized subsample. The sample is further restricted to borrowers initially assigned on day 2 to callers who specialize in debts that are in their first five days past due.[20] In untabulated results, we find that differences in overdue payment amount, remaining principal, internal credit score, gender, age, or attainment of a bachelor's degree measured on day 2 after the due date between borrowers assigned to junior versus senior callers have t-statistics whose magnitude is 1.47 or less when month fixed effects are controlled for.

[20] Callers in other specialties are also assigned these cases if the caseload is high. Restricting to day 2-5 specialists creates a more uniform sample over time.

Table 10 shows that borrowers initially assigned to junior callers repay 1.1 percentage points less normalized NPV on day 2, but the gap disappears in the following days as the debts are rotated to other callers on each day. Again, there is no permanent damage associated with a less productive initial human contact. In Panel B, we further split the group of junior callers by their work experience and find that, although underperformance is larger and persists for a few more days among workers with no more than one month of experience, the productivity gap is mostly offset in the long run.

These results indicate that there is something uniquely damaging about being contacted by an AI caller. However, the evidence is only suggestive because the less productive initial human contacts we test are still much more effective than AI callers. We cannot rule out that calls from a human who is just as ineffective as an AI caller would result in similar long-term damage.

6 Impacts of AI on Human Callers

6.1 Impact of AI upgrades on human caller productivity

The AI caller software was upgraded several times during our sample period, which allows us to examine how improvements in AI productivity affects human callers' productivity. This question is particularly important in light of the current rapid development of AI technology.

Figure 8 illustrates the AI upgrade process by showing the fractions of cases assigned to different versions of AI callers each month in our sample period. We study six versions of AI callers. We call the first version in our sample period "V1," although this is not the very first version of the AI deployed by the company in 2018. Subsequent versions are labeled "V2" to "V6," according to their order of introduction.

As Figure 8 shows, the company introduced new versions of AI callers progressively. V1 was the predominant version used at the beginning of the sample period in April 2021, but the company was already simultaneously deploying V2. The first cases assigned to V3 were in May 2021, and V2 was phased out after June 2021. V4 was introduced in September 2021 and took over all cases starting in November 2021. The company began testing V5 in January 2022 and retired V4 after February 2022. The final version in our sample period, V6, began testing in June 2022 and took over all cases starting in October 2022.

Since assignment to different active versions of AI callers was random within each time period, measuring their relative productivity is straightforward. In the following analyses, we restrict our sample to the completely randomized subsample. Figure 9 shows the time series of average collected NPVs on day 2 and over the first five days past due for each AI caller version, along with the corresponding collections of human callers. We see that the gap between AI and human callers remains wide over time, despite multiple AI upgrades.

Table 11 formally tests the improvement of AI performance, comparing cumulative NPV collected over two- to ten-day horizons between pairs of contemporaneously active AI callers. Recall that in the completely randomized subsample, the AI caller only works from days 2-5, so collections after day 5 reflect human caller effort. For each NPV horizon, the test is implemented by regressing collected NPV onto a set of indicators for which AI version was assigned the case and calendar month dummies. The results show that the most salient improvement occurred when upgrading from V1 to V4. The increase in collected NPV is 0.030 over the first five days, which is 8% of the average five-day NPV collected for all AI-assigned cases in the completely randomized subsample, shown in the last row. On day 6, however, human callers take over, immediately closing the cumulative gap to a statistically insignificant 0.012 between borrowers initially called by V4 and V1. This means that human callers taking over from V4 collect 0.030 – 0.012 = 0.019 less NPV on day 6 than humans taking over from V1.[21]

The day 6 finding suggests that the more effective AI callers are in the first five days, the tougher the cases that are left for human callers on day 6, and the less human callers can collect. Across the five AI upgrades, there is no statistically robust evidence that the total amount collected by humans partnering with AI expands. This result is consistent with a displacement effect of AI callers.

In untabulated analysis, we test whether the long-run damage from AI contact through day 5 followed by human contact is mitigated when moving from V1 to V4. We find no statistically significant differences in NPV collected through day 180 or 360, nor in the likelihood of missing the next twelve monthly payments due. There is no tendency for the point estimates to hint at smaller damage from V4 than V1.

[21] The company says that the AI improvements were concentrated in the speech recognition and language understanding algorithms. Appendix Table B4 compares several phone call outcomes across different versions of AI. From V1 to V4, the average phone call duration falls, but there are also fewer calls shorter than 10 seconds. Given that most long AI calls consist of repetitious conversations, this result indicates a potential increase in the information efficiency of the conversations. In addition, V4 is better at extracting promises to repay from borrowers. Finally, V4 has fewer calls classified as ineffective, which includes calls with only silence after picking up, voicemail or auto-reply voice messages, and conversations that the AI cannot understand or cannot classify into pre-specified scenarios.

6.2 AI productivity net of labor cost savings

AI callers perform significantly worse than human callers. On the other hand, AI callers have almost zero marginal costs when making phone calls. Therefore, to fully evaluate how well AI may replace human callers, we need to subtract labor costs in our NPV calculations.

Aggregate in-house caller salary consists of two parts. One part is fixed, depending only on the total number of callers and unrelated to how much money they collect. The other part is variable, a function of the total amount of money that callers collect. Although there is nonlinearity in the salary schemes, for a back-of-envelope calculation, we approximate the structure of the labor costs as follows.

Dividing total fixed salary paid to in-house callers specializing in a given stage of collection each month by the total number of minutes of phone calls made by these callers in that month, and computing the time-series average of this monthly ratio, we estimate that the fixed cost per phone call minute is 1.157, 1.313, 1.110, and 0 yuan for calls on days 2-10, 11-25, 26-59, and 60-84, respectively.[22] Similarly dividing total variable salary by total yuan collected each month and computing the time-series average of this monthly ratio, we estimate that the variable cost per yuan collected is 0.0051, 0.0186, 0.0274, and 0.0743 yuan for calls on days 2-10, 11-25, 26-59, and 60-84, respectively.

Nearly all cases beyond day 84 are handled by third-party collectors. Third-party collectors are paid solely by commissions and performance bonuses. In the third quarter of 2023, payments to third-party collectors as a percent of the amount they collected were 28.8% and 16.8% for collections in days 91-180 and 181-360, respectively. We do not have information on third-party payments for collections in earlier stages, so we use total in-house compensation as a percent of in-house collections for our third-party cost calculation: 1.05%, 3.59%, 5.63%, and 10.27% for collections in days 2-10, 11-25, 31-60, and 61-90, respectively.

Figure 10 shows the average differences in collected NPV between AI and human callers as a function of days past due after adjusting for caller salary. In the NPV calculation, we subtract out labor costs on the same day that the associated collection effort occurs. The per-call-minute fixed cost is converted to a borrower-day fixed cost by multiplying the per-call-minute cost by the average phone call length per still-delinquent borrower on the corresponding day after delinquency

[22] The fixed salary was non-zero for calls to borrowers in days 60-84 prior to September 2021. We use the post-September 2021 salary scheme throughout our calculation for simplicity.

in the relevant subsample. Panel (a) uses the RD subsample, and Panels (b) and (c) use the completely randomized subsample equally weighted or weighted by the overdue payment amount on day 2, respectively. The estimation methods are the same as what we use for the corresponding subsamples in Figure 4 and Figure 6.

After accounting for direct labor costs, the collected NPV gaps between AI and human callers become narrower. For small cases around 300 yuan of remaining principal, the collected NPV difference between AI and humans remains negative at all horizons but becomes statistically insignificant after about eight months of collection effort. In the completely randomized subsample, where the case sizes are larger, the adjusted productivity gaps become smaller and eventually insignificant when cases are equal-weighted, but AI remains statistically significantly less cost-effective than humans even in the long run when cases are value-weighted (NPV difference at day 360 = –0.0039, p = 0.033). Because calling one borrower incurs approximately the same fixed labor cost regardless of the debt's size, the return to human labor is relatively low on smaller cases.[23]

Importantly, we only have considered direct labor costs, i.e., salary paid to human callers. To hire and manage more than 2,000 callers, the company also needs to spend money on many indirect costs, such as worker recruitment, training, management, pension funds, etc. On the other hand, we also do not include in our calculation the cost of developing and improving the AI software. The cost-effectiveness of AI will also depend on the wage levels in the local labor market.

7 Conclusion

An important task in many service and managerial jobs is to persuade another human to perform a personally costly action. In this paper, we study how well AI does in one instance of such a task: persuading borrowers to pay their delinquent debt. We find that currently, AI is significantly worse than humans at getting borrowers to repay. AI appears to have less ability to extract verbal promises to repay and create a sense of obligation to keep those promises, perhaps because AI carries less moral weight. AI additionally seems less able to communicate with borrowers, engage their attention, and/or handle complex situations.

[23] Human callers do make somewhat shorter calls to borrowers with small balances, reducing the fixed labor cost attributed to calling those borrowers. See Online Appendix Figure .

Losses from using AI can be substantially mitigated if human callers take over cases from AI after a few days. Nonetheless, even in this collaborative arrangement, borrowers who were initially called by AI have paid less after a year than borrowers who were always called by humans. The alienation created by AI contact extends beyond the initial late payment in question; AI-contacted borrowers are more likely to miss each of their next 12 monthly payments. Thus, AI callers create enduring damage to the company's relationship with its borrowers.

Of course, AI technology will continue to advance. Although AI-powered robots will probably remain easily distinguishable from humans in face-to-face encounters for a long time—meaning AI will continue to be less effective at arm-twisting in in-person settings—we may soon reach a point where it is difficult for people to discern whether a caller on the phone is a human or an AI. At that point, people may assume that *all* callers are AIs. Newman (2019) presents a thought experiment involving an authentic Picasso painting and a perceptually identical forgery. It is obvious that people value the forgery less than the authentic painting, and if people believe the authentic Picasso is a forgery, they will react much less positively to it. Similarly, if an important driver of the negative reaction to AI is the belief that it is not a human rather than the perceivable experience of talking to an AI, then future call recipients may become less cooperative with all callers, regardless of whether the caller is actually human.

References

Acemoglu, Daron, and Pascual Restrepo. 2020. "AIs and jobs: Evidence from US labor markets." *Journal of Political Economy*, 128(6): 2188-2244.

Acemoglu, Daron, and Pascual Restrepo. 2022. "Tasks, automation, and the rise in us wage inequality." *Econometrica*, 90(5): 1973-2016.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. 2019. "Artificial intelligence: the ambiguous labor market impact of automating prediction." *Journal of Economic Perspectives*, 33(2): 31-50.

Aydin, Deniz. 2022. "Consumption response to credit expansions: Evidence from experimental assignment of 45,307 credit lines." *American Economic Review*, 112(1): 1-40.

Brynjolfsson, Erik, Danielle Li, and Lindsey R. Raymond. 2023. "Generative AI at work." *NBER Working Paper* No. w31161.

Brynjolfsson, Erik, and Tom Mitchell. 2017. "What can machine learning do? Workforce implications." *Science*, 358(6370): 1530-1534.

Burstyn, Leonardo, Stefano Fiorin, Daniel Gottlieb, and Martin Kanz. 2019. "Moral incentives in credit card debt repayment: Evidence from a field experiment." *Journal of Political Economy*, 127(4): 1641-1683.

Cao, Sean, Wei Jiang, Junbo L. Wang, and Baozhong Yang. 2024. "From man vs. machine to man+ machine: The art and AI of stock analyses." *Journal of Financial Economics*, 160: 103910.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2019. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and partnership." *Econometrica*, 74(6): 1579-1601.

Cheng, Ing-Haw, Felipe Severino, and Richard R. Townsend. 2021. "How do consumers fare when dealing with debt collectors? Evidence from out-of-court settlements." *Review of Financial Studies*, 34(4): 1617-1660.

Cohn, Alain, Tobias Gesche, and Michel André Maréchal. 2022. "Honesty in the digital age." *Management Science*, 68(2): 827-845.

Drozd, Lukasz A., and Ricardo Serrano-Padial. 2017. "Modeling the revolving revolution: the debt collection channel." *American Economic Review*, 107(3): 897-930.

Ellingsen, Tore, and Magnus Johannesson. 2004. "Promises, threats and fairness." *Econnomic Journal*, 114(495): 397-420.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock. 2024. "GPTs are GPTs: Labor market impact potential of LLMs." *Science*, 384(6702): 1306-1308.

Erel, Isil, Léa H. Stern, Chenhao Tan, and Michael S. Weisbach. 2021. "Selecting directors using machine learning." *Review of Financial Studies*, 34(7): 3226-3264.

Fedaseyeu, Viktar. 2020. "Debt collection agencies and the supply of consumer credit." *Journal of Financial Economics*, 138(1): 193-221.

Fedaseyeu, Viktar, and Robert M. Hunt. 2015. "The economics of debt collection: Enforcement of consumer credit contracts." *FRB of Philadelphia Working Paper* No. 15-43.

Felten, Edward W., Manav Raj, and Robert Seamans. 2020. "The occupational impact of artificial intelligence: Labor, skills, and polarization." *SSRN Working Paper* No.3368605.

Gao, Zihan, and Jiepu Jiang. 2021. "Evaluating human-AI hybrid conversational systems with chatbot message suggestions." In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*: 534-544.

Glikson, Ella, and Anita Williams Woolley. 2020. "Human trust in artificial intelligence: Review of empirical research." *Academy of Management Annals* 14(2): 627-660.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales. 2013. "The determinants of attitudes toward strategic default on mortgages." *Journal of Finance*, 68(4): 1473-1515.

He, Simin, Theo Offerman, and Jeroen Van De Ven. 2017. "The sources of the communication gap." *Management Science*, 63(9): 2832-2846.

Karlan, Dean, Melanie Morten, and Jonathan Zinman. 2015. "A personal touch in text messaging can improve microloan repayment." *Behavioral Science & Policy*, 1(2): 25-31.

Kim, Tae Woo, Hyejin Lee, Michelle Yoosun Kim, SunAh Kim, and Adam Duhachek, 2023. "AI increases unethical consumer behavior due to reduced anticipatory guilt." *Journal of the Academy of Marketing Science*, 51: 785-801.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human decisions and machine predictions." *Quarterly Journal of Economics*, 133(1): 237-293.

LaMothe, Ethan, and Donna Bobek. 2020. "Are individuals more willing to lie to a computer or a human? Evidence from a tax compliance setting." *Journal of Business Ethics*, 167: 157-180.

Laudenbach, Christine, and Stephan Siegel. 2023. "Personal communication in an automated world: Evidence from loan repayments." *Journal of Finance*, forthcoming.

Luo, Xueming, Siliang Tong, Zheng Fang, Zhe Qu. 2019. "Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases." *Marketing Science*, 38(6): 937-947.

Luo, Xueming, Marco Shaojun Qin, Zheng Fang, and Zhe Qu. 2021. "Artificial intelligence coaches for sales agents: Caveats and solutions." *Journal of Marketing*, 85(2): 14-32.

Melo, De Celso, Stacy Marsella, and Jonathan Gratch. 2016. "People do not feel guilty about exploiting machines." *ACM Transactions on Computer-Human Interaction,* 23: 1–17.

Michaels, Guy, Ashwini Natraj, and John Van Reenen. 2014. "Has ICT polarized skill demand? Evidence from eleven countries over twenty-five years." *Review of Economics and Statistics*, 96(1): 60-77.

Newman, George. 2019. "The psychology of authenticity." *Review of General Psychology*, 23(1), 8-18.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." *Science*, 381(6654): 187-192.

Petisca, Sofia, Ana Paiva, and Francisco Esteves. 2020. "Perceptions of people's dishonesty towards AIs." *International Conference on Social AIics*: 132–143. Springer.

Roll, Stephen P., and Stephanie Moulton. 2019. "The impact of automated reminders on credit outcomes: Results from an experimental pilot program." *Journal of Consumer Affairs*, 53(4): 1693-1724.

Silva, Gisele S., Rohan Khera, and Lee H. Schwamm. 2024. "Reviewer experience detecting and judging human versus artificial intelligence content: The *Stroke* journal essay contest." *Stroke* 55: 2573-2578.

Tan, Tom Fangyun, and Serguei Netessine. 2020. "At your service on the table: Impact of tabletop technology on restaurant performance." *Management Science*, 66 (10): 4496-4515.

Vanberg, Christoph. 2008. "Why do people keep their promises? An experimental test of two explanations." *Econometrica*, 76(6): 1467-1480.

World Economic Forum. 2020. *The Future of Jobs Report 2020*. Switzerland.

World Economic Forum. 2023. *The Future of Jobs Report 2023*. Switzerland.

Zhou, Yijun. 2024. "Artificial intelligence and debt collection: Evidence from a field experiment." SSRN Working paper No.4905228.

Figures

Figure 1. Case assignment between AI and human callers on day 2 past due.

This figure shows how cases are assigned between AI and human callers. "Almost always AI" means that more than 95% of cases are always handled by AI callers over the life of the cases, and less than 5% of cases are assigned to human callers after day 25. For conditionally randomized borrowers, the probability of being assigned to a human caller and when humans take over from AI depend on case characteristics and the calendar date.

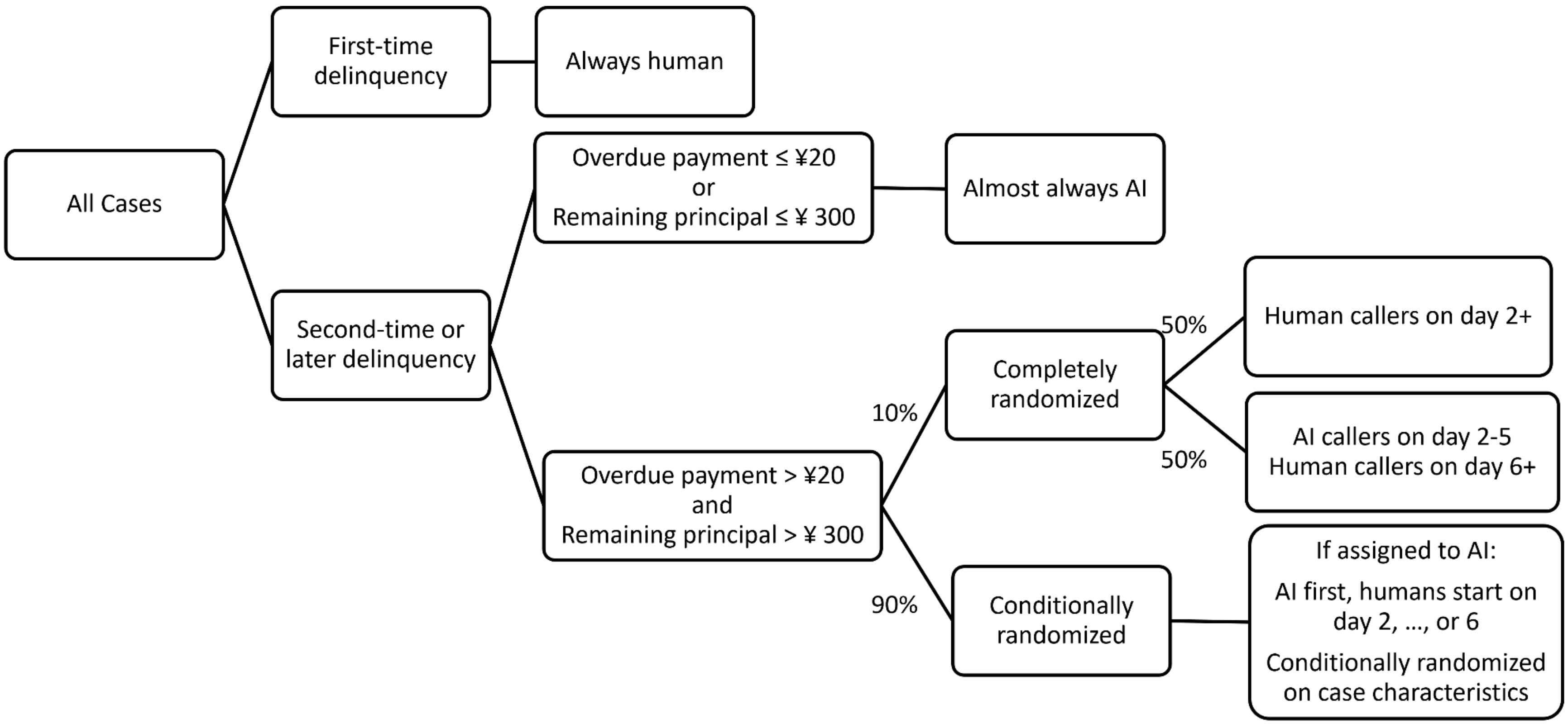
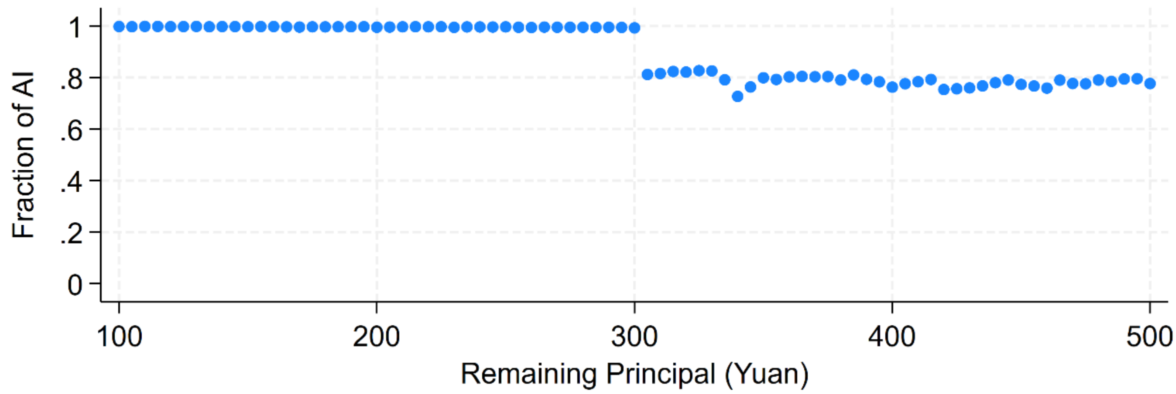


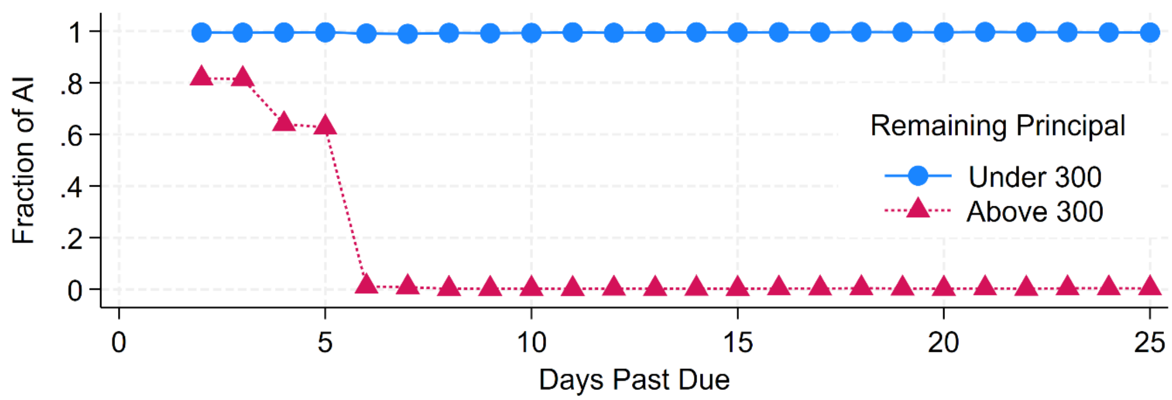
Figure 2. Fraction of cases assigned to AI callers around the 300-yuan remaining principal threshold.

Panel (a) shows the fraction of cases assigned to AI callers on day 2 past due as a function of principal remaining on day 2. Panel (b) shows the AI fractions for cases in the (295, 300] and (300, 305] yuan day-2 principal remaining intervals on days 2-25 past due. Panel (c) extends the horizon of panel (b) to day 360.

(a) Fraction of AI cases by remaining principal on day 2 past due



(b) Fraction of AI cases just below and above the 300-yuan threshold (days 2-25 past due)



(c) Fraction of AI cases just below and above the 300-yuan threshold (days 2-360 past due)

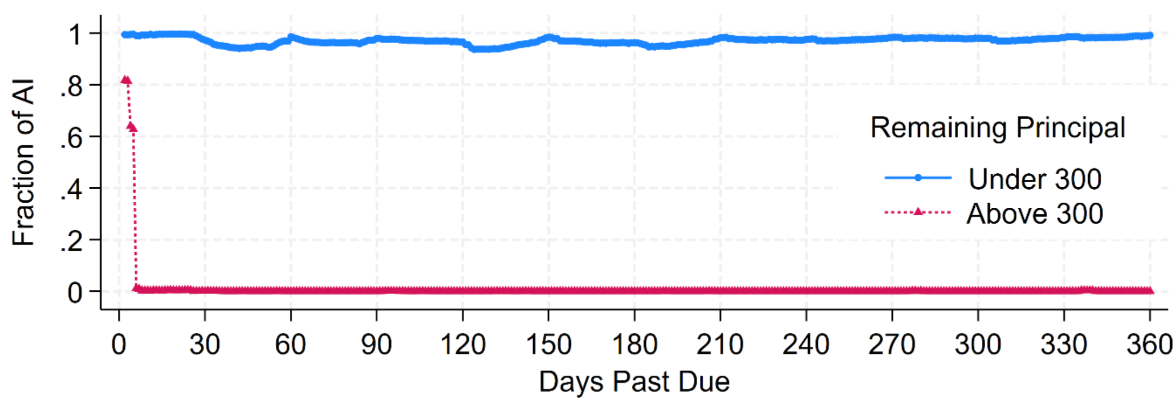
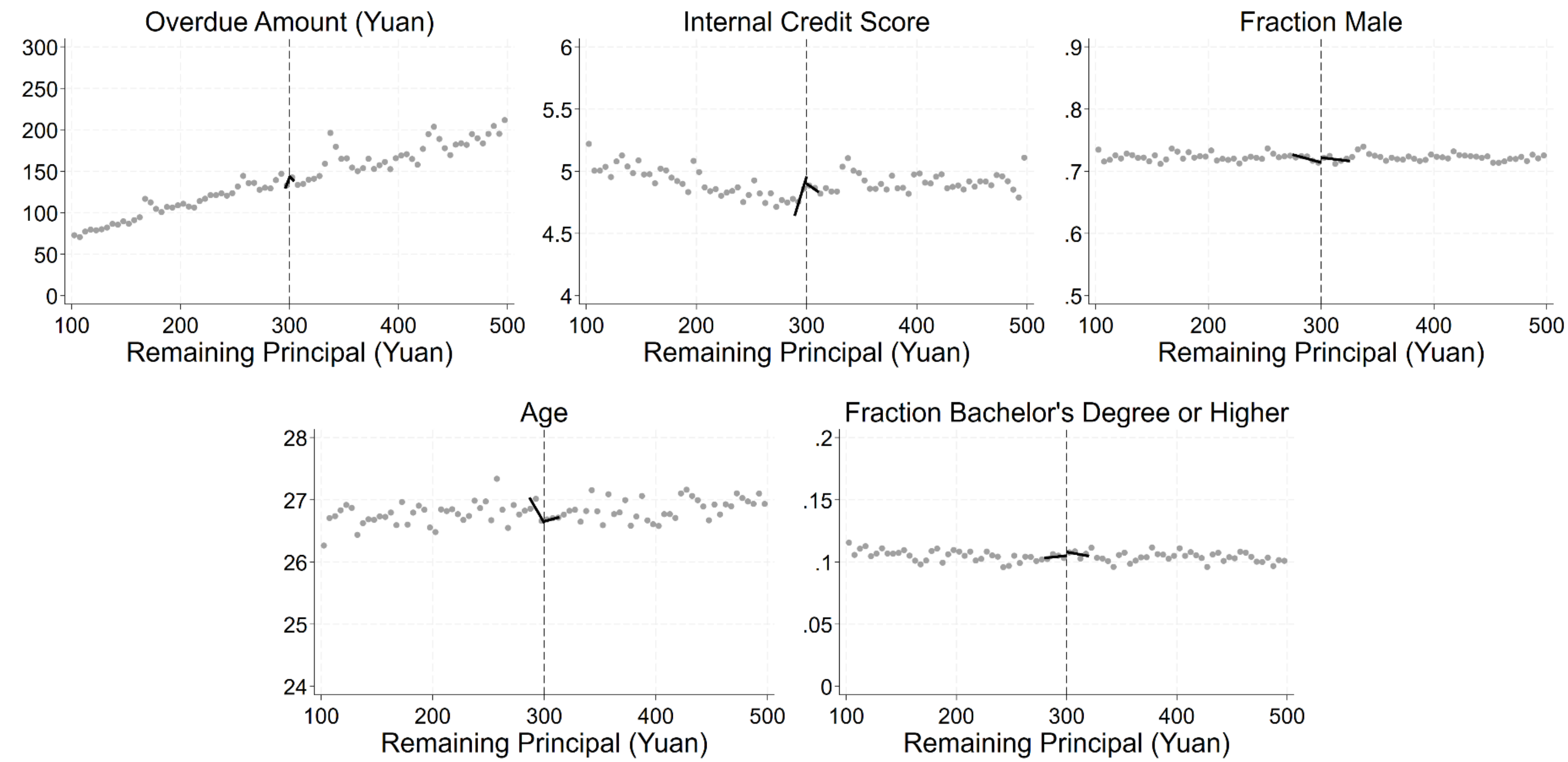


Figure 3. Loan and borrower characteristics and collected NPVs around the 300-yuan remaining principal cutoff.

Panel (a) plots average loan and borrower characteristics measured on day 2 past due as a function of remaining principal on day 2 past due. Panel (b) plots average cumulative NPVs of payments collected up until day 2, 5, 10, 30, 90, or 360 past due, divided by the balance due on day 2, as a function of remaining principal on day 2 past due. The grey dots are binned averages, and the black lines are local linear fits within the regression discontinuity estimation bandwidths on each side. Coverage-error-rate-optimal bandwidths are used in Panel (a) and mean-squared-error-optimal bandwidths in Panel (b).

(a) Loan characteristics



(b) Cumulative collected NPV

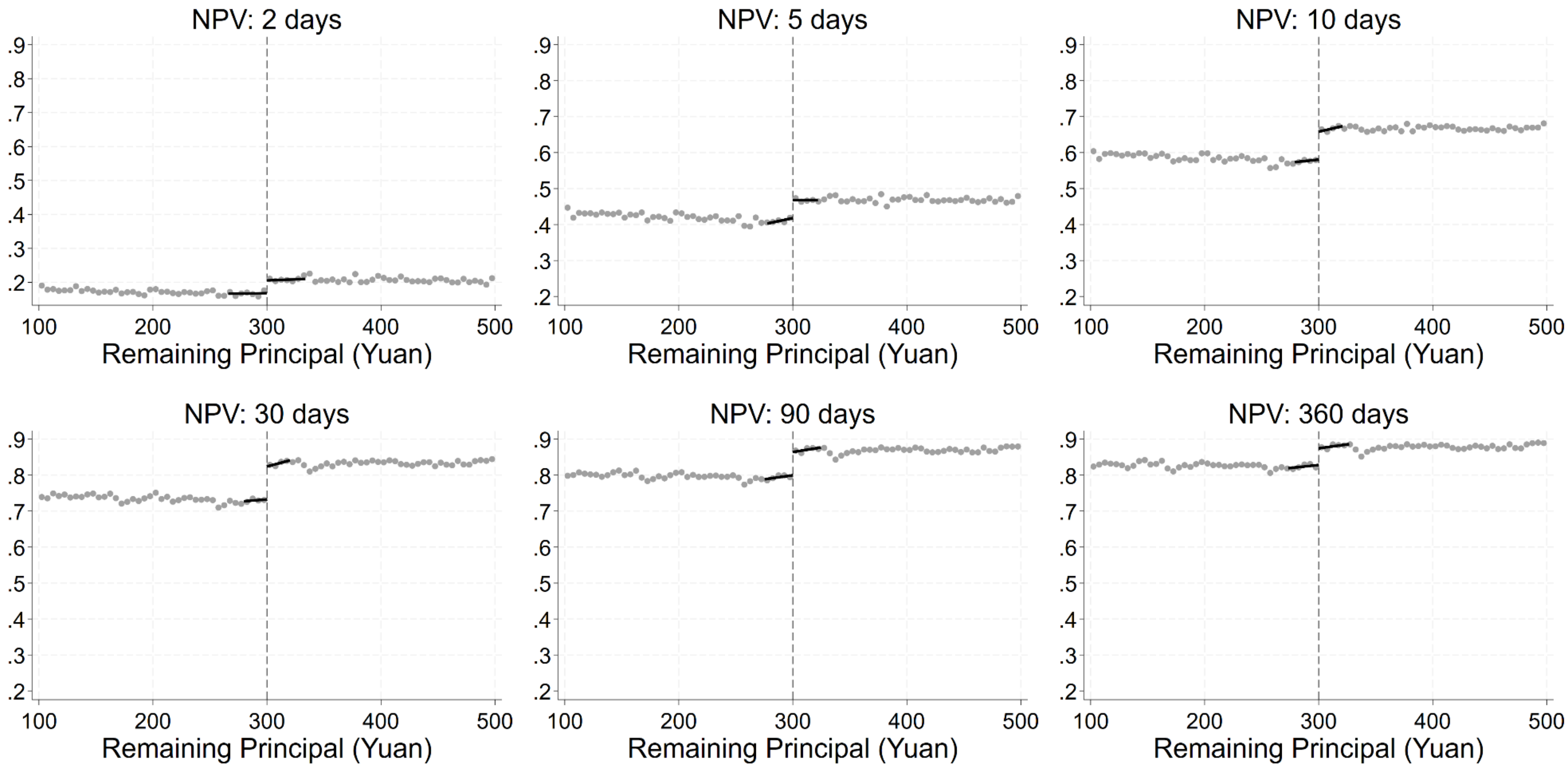


Figure 4. AI minus human caller cumulative collected NPV by horizon – small cases RD.

This figure shows the difference in average cumulative NPVs of payments collected divided by balance due on day 2 between AI and human callers as a function of days past due. The differences are estimated by regression discontinuity around the 300-yuan remaining principal threshold for permanent AI treatment. The bars indicate 95% robust bias-corrected regression discontinuity confidence intervals. Differences are plotted every three days before day 60 and every 10 days after day 60.

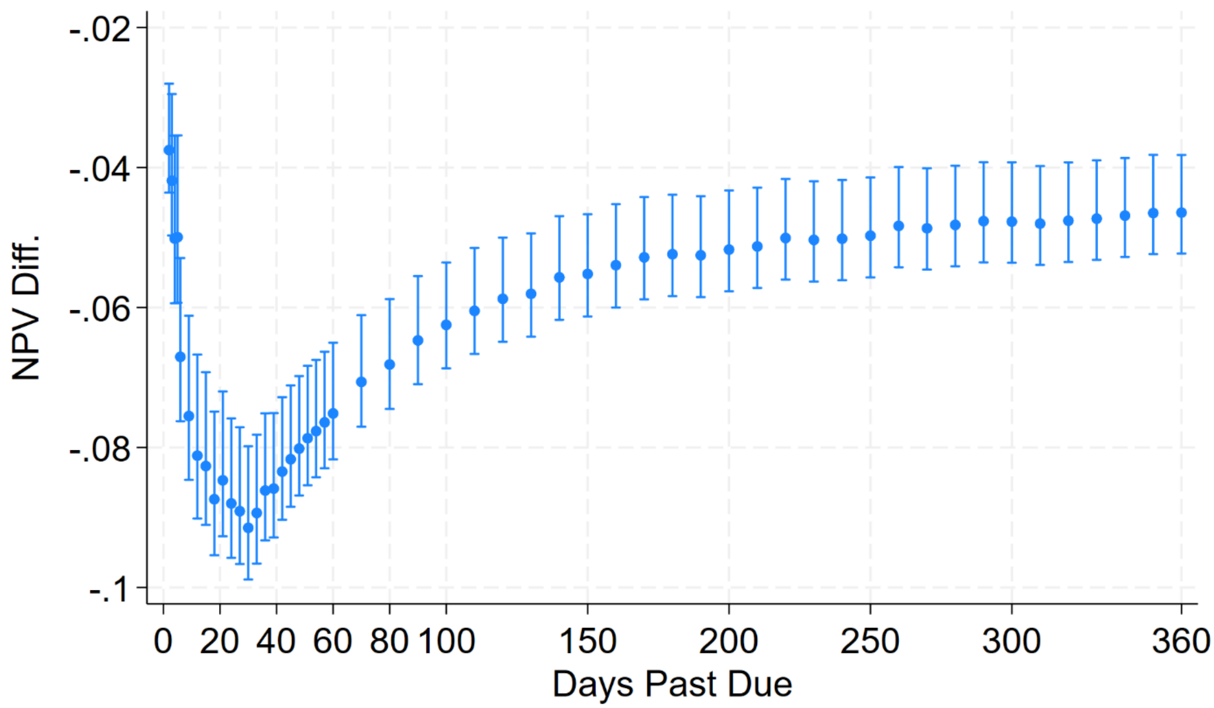


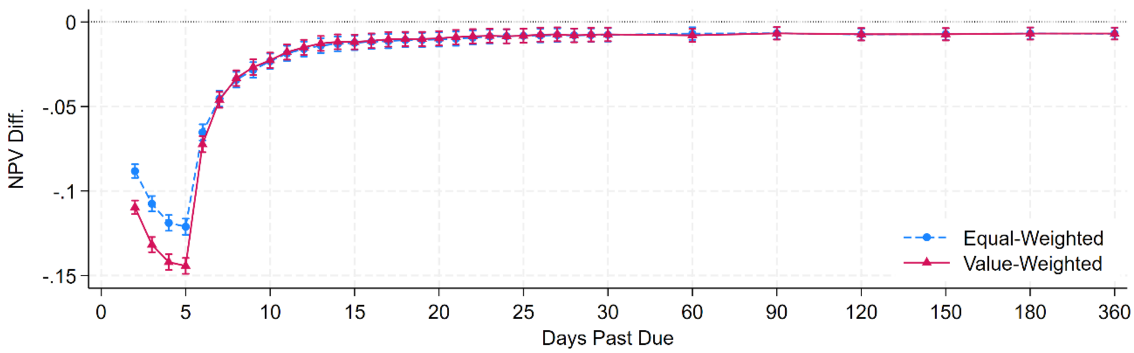
Figure 5. AI minus human caller cumulative collected NPV by horizon and internal credit score – small cases RD.

This figure shows the differences in average cumulative payment NPVs collected divided by balance due on day 2 between AI and human callers as a function of days past due, separately for borrowers in three groups of internal credit scores measured on day 2. The differences are estimated by regression discontinuity around the 300-yuan remaining principal threshold for permanent AI treatment. “Low”, “Med”, and “High” refer to cases with internal credit scores in deciles 1-3, 4-7, and 8-10, respectively. Differences are plotted every three days before day 60 and every 10 days after day 60.

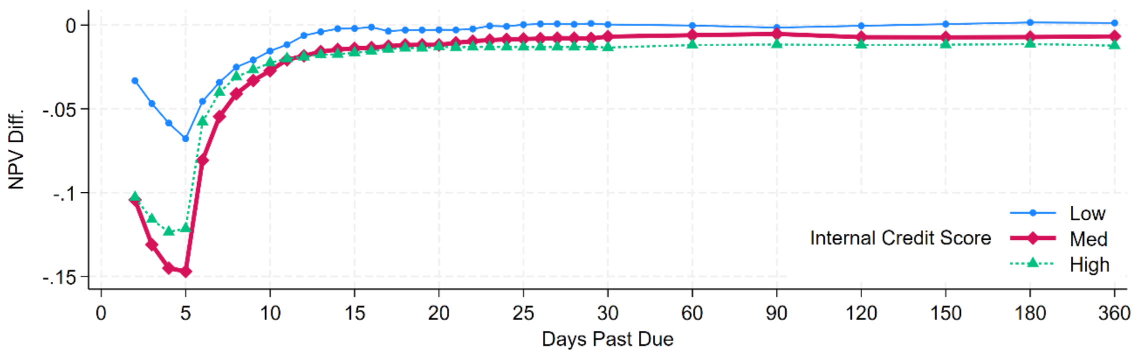
Figure 6. AI + human minus always-human cumulative collected NPV by horizon – Completely randomized subsample.

This figure shows the difference in average cumulative NPVs of payments collected divided by balance due on day 2 between the AI + human and always-human groups as a function of days past due, using the completely randomized subsample. The differences control for calendar month of day 2 fixed effects. The bars indicate 95% confidence intervals. Differences are plotted daily before day 30 and every 30 days afterwards. Panel (a) pools all cases together. Panels (b) and (c) split cases by internal credit score and overdue payment size, respectively, measured on day 2. In Panel (b), "Low", "Med", and "High" refer to internal credit scores in deciles 1-3, 4-7, and 8-10, respectively.

(a) All cases.



(b) By internal credit score.



(c) By loan size.

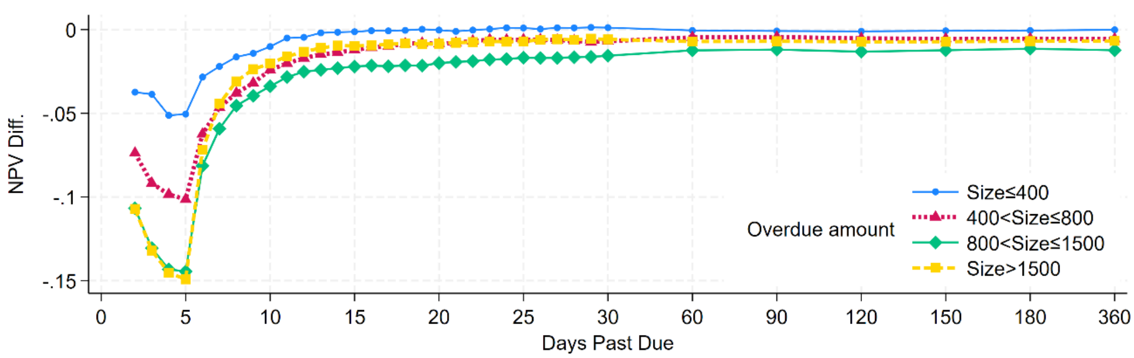


Figure 7. Percent of calls made to emergency contacts – Completely randomized subsample.

The sample is split into calls made before April 2022 and those made from April 2022 onwards.

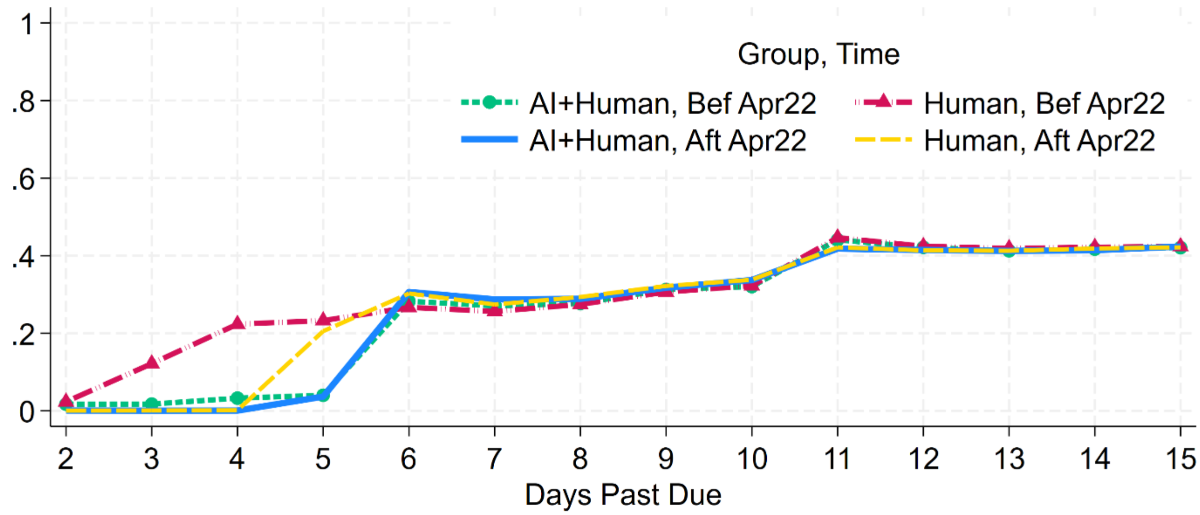


Figure 8. Deployment of different versions of AI callers over time.

This figure shows the fraction of cases assigned to each version of AI callers during each month of our sample period. The first version in our sample period is labeled "V1," but it is not the very first version of AI caller used by the company. The fractions are calculated within the completely randomized subsample.

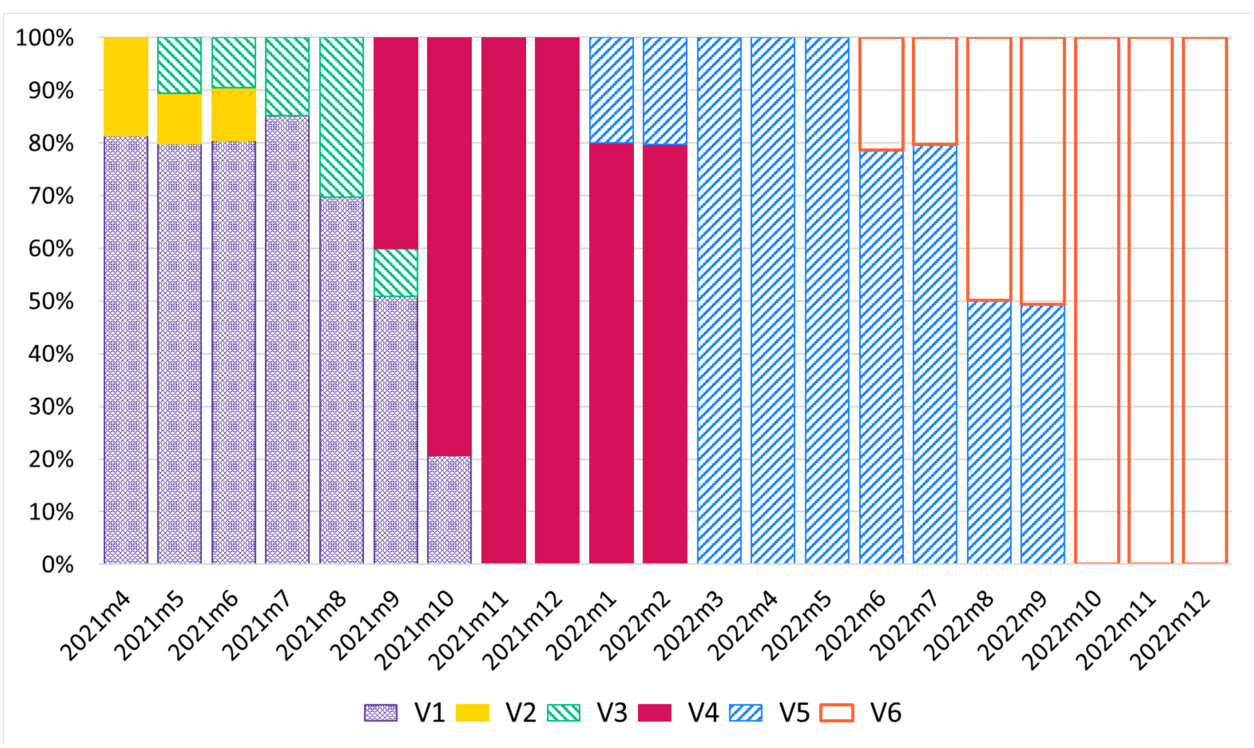
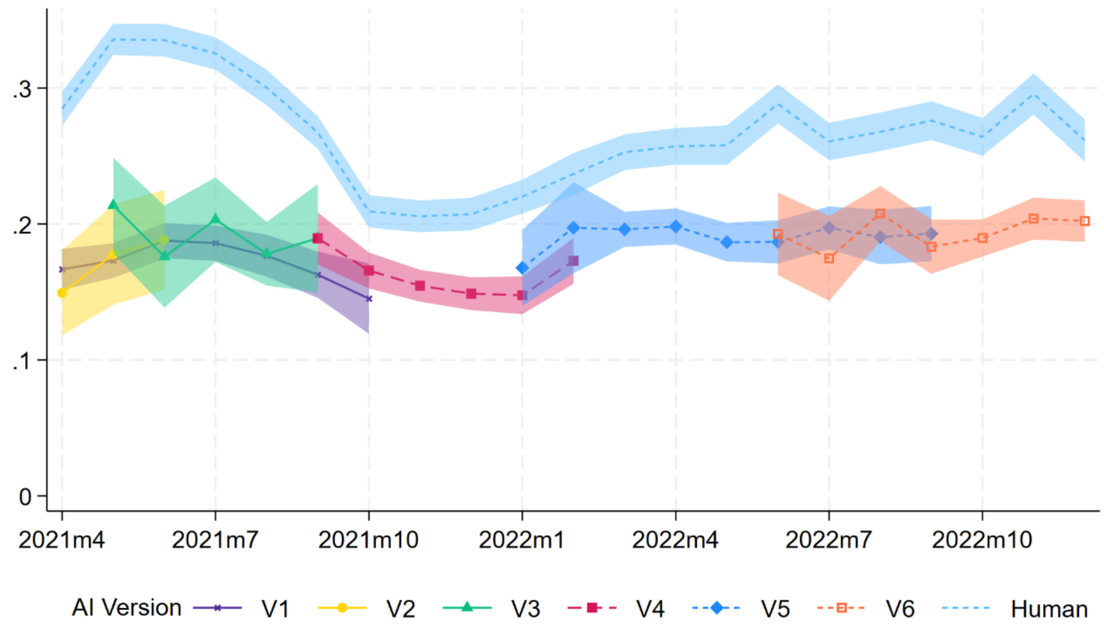


Figure 9. Performance of different versions of AI callers over time.

This figure shows the monthly time-series of performance of different versions of AI callers and human callers, measured by the average cumulative collected NPV on day 2 past due (Panel (a)) or through day 5 past due (Panel (b)). The shaded areas represent 95% confidence intervals.

(a) NPV on day 2



(b) NPV over first 5 days

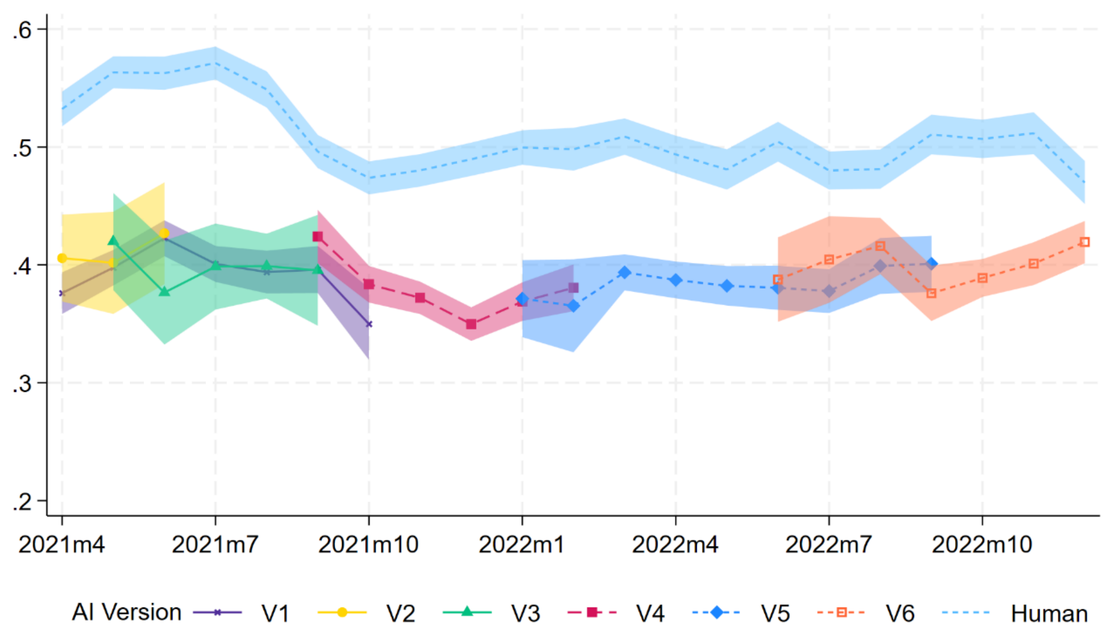
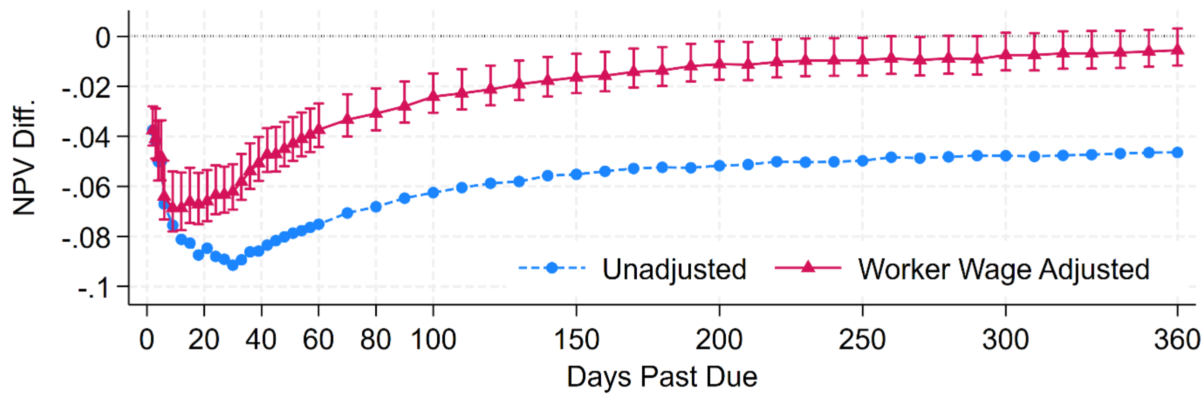


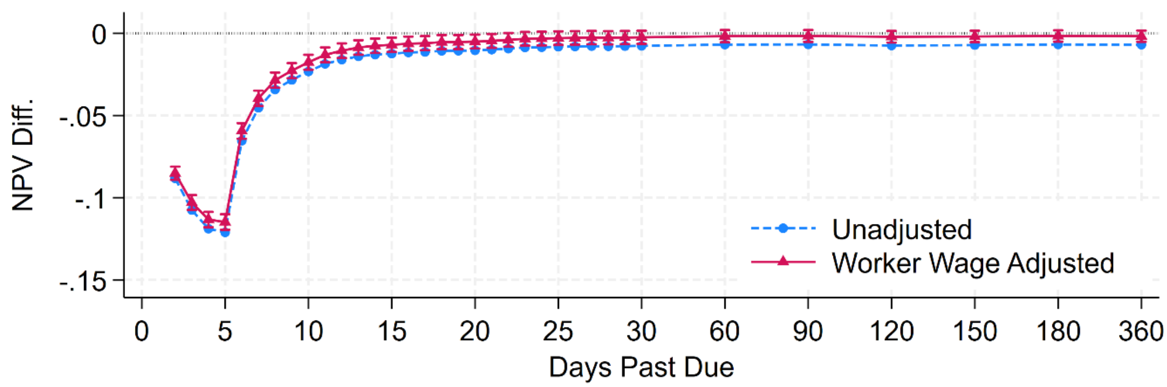
Figure 10. Collected NPV differences between AI and human callers over time, adjusted for labor costs.

This figure shows the difference in average cumulative NPVs of payments collected net of labor costs divided by balance due on day 2 between AI and human callers as a function of days past due. The bars indicate 95% confidence intervals. Panel (a) shows results from small cases using a regression discontinuity design. Panels (b) and (c) use the completely randomized subsample, with the former equally weighting all cases and the latter weighting cases by the overdue payment amount on day 2.

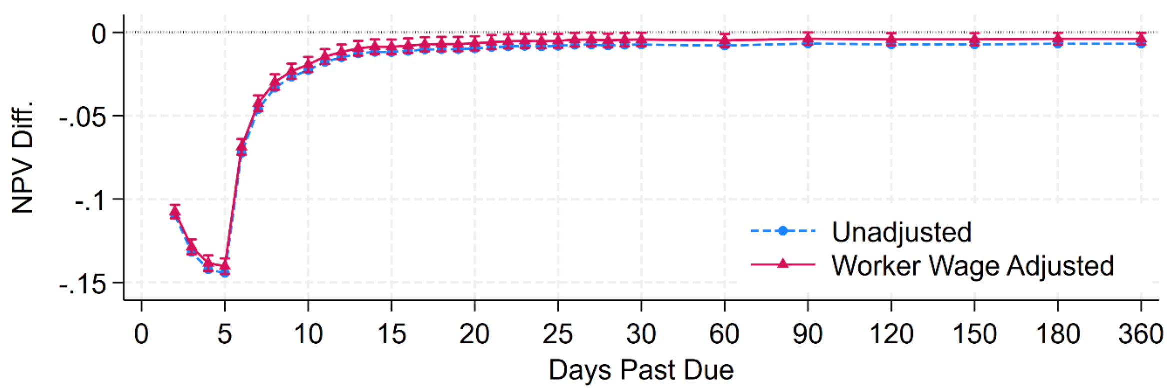
(a) Small cases – RDD.



(b) Completely randomized subsample, equal-weighted.



(c) Completely randomized subsample, value-weighted.



Tables

Table 1. Sample scripts of AI caller.

This table presents sample scripts that the AI caller typically uses in a phone call for debt collection. The third stage may include multiple rounds of questions and responses.

Stage	Stage Description	Sample Scripts
1	Greeting. Confirm the identity of the borrower.	AI: Hello, is this [borrower's full name]?
	Case 1.1. It's the borrower.	Borrower: Yes. [AI moves to Stage 2]
	Case 1.2. Recipient claims to not be the borrower.	Borrower: No, you have the wrong number. AI: Then have you ever borrowed money from [Lending Company]?
	Case 1.2.2. Recipient has borrowed from the company.	Borrower: Yes. [AI moves to Stage 2]
	Case 1.2.3. Recipient has not borrowed from the company.	Borrower: No. AI: Okay, this is [Lending Company]. This phone number is related to one of our delinquent borrowers. If you are the owner of this phone number, could you please contact our customer service and resolve the mistake? Sorry for disturbing you and goodbye! [or] AI: Then do you know [borrower's full name]? If so, could you let him/her know that he/she has an overdue bill with [Lending Company], and that he/she should repay it as soon as possible to avoid any negative consequences? Sorry to interrupt and goodbye!
	Case 1.3. The recipient claims to be the borrower's social contact.	Borrower: I'm his/her sister. AI: This is [Lending Company]. [Borrower's full name] has a bill with us that he/she probably forgot to repay, and it is now severely overdue. Could you convey the message to him/her? Borrower: Okay. AI: Okay. Thank you. We will check his/her account balance by 5 P.M. today. Please ask him/her to repay as soon as possible to avoid negative impacts on her credit records. Sorry for disturbing you. Goodbye.

2	Inform the borrower about the delinquent debt. Ask them to repay within two hours.	AI: This is [Lending Company]. You have a bill of y yuan that has been overdue for z days. Could you please repay it within two hours?
3	*Case 3.1*. Borrower agrees to repay within two hours.	Borrower: Yes, I will repay the debt. AI: Continuing to be overdue will negatively affect your credit. You're sure you can repay within two hours, right? Borrower: Yes. AI: All right! I have recorded it in our system. I have temporarily suspended the debt collection process for you. Please repay the debt as soon as possible. If we do not receive your payment on time, we will contact you again / your overdue record will be automatically uploaded to the central bank's credit reporting agency / your bill will be transferred to specialized debt collection agencies immediately / we may contact your emergency contacts for further information about your financial conditions / [other negative consequences].
	Case 3.2. Borrower wants to repay later than two hours or today.	Borrower: No, I cannot do it today. How about tomorrow? AI: Your loans are related to your credit records. If you continue to delay, your case will enter the debt collection process. Can you repay *today*? Borrower: Tomorrow is the best I can do. AI: Okay. We have noted it down. I have temporarily suspended the debt collection process for today. / No late fees will be charged for today. But if you continue to be in default, your credit records will be negatively affected / [other negative consequences].
	Case 3.3. Borrower refuses to repay and sometimes explains the reason.	Borrower: I cannot repay today because I have no money to repay / I am very busy today / Sorry I forgot / I'm now in the hospital. AI: You can always find a way to raise money. For example, you can borrow from your family members and friends. / Even if you are busy, I believe you could still find just a few minutes to repay the debt. / My guess is also that you just forgot to repay. / Sorry to hear that but since your debt has been delinquent, could you repay it now? Continuing to default will [potential negative consequences]. Can you repay the debt by the end of today?
	Case 3.4. Borrower denies having debts with the company or wants to verify the name of the company.	Borrower: I don't know [Lending Company]. / I have never borrowed money from your company. / What company are you with? AI: This is [Lending Company] in [city name]. The name is spelled as [spelling]. You have an overdue debt with us.

	Case 3.5. Borrower claims to have repaid the debt.	Borrower: I have already paid back the debt this morning, haven't you received the money yet? AI: But there is still *y* yuan on your balance according to the records in our system. We will double-check our records. You can also log into our mobile app to verify whether you've already repaid the entire balance.
	Case 3.6. Borrower claims to have set up auto-payment.	Borrower: I have set up auto-payment. AI: Okay. We will try to charge *y* yuan from your linked bank account shortly. Please make sure that you have sufficient balances in your bank account.
	Case 3.7. Borrower asks for additional information, including days past due, overdue balance amount, late fees, due date, payment methods, etc.	Borrower: Do I need to pay any late fees? AI: Late fees include overdue interest and principal, credit evaluation fees, guarantee fees, and so on. Details can be found in our app or contact customer service. Borrower: How long has it been overdue? AI: You have been *z* days past due. We have sent you several text messages before.
4	When the borrower has no more questions, the borrower's questions do not belong to the above cases, or the AI cannot understand the borrower's response.	AI: Okay. Please be advised that you will be responsible for any negative consequences of default [or emphasize one of the negative consequences.] If you have any other questions, feel free to contact our customer service. Bye!

Table 2. Summary statistics of delinquent loans.

This table reports summary statistics for the full sample of delinquent loans and two different subsamples used in our analyses. Variables are measured on day 2 past due. Panel A shows the mean, standard deviation, selected percentiles, and observation counts for the full sample. Panel B shows statistics for delinquent loans with remaining principal between 100 yuan and 500 yuan. Panel C shows statistics for the completely randomized subsample, which is restricted to borrowers in their second delinquency.

Panel A. Full sample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	# Obs.
Overdue amount (yuan)	1,128.1	1,822.4	0.01	14.7	316.0	653.5	1,304.6	7,688.8	808,666.7	22,122,179
Remaining principal (yuan)	6,474.0	7,330.0	0.01	48.6	1,792.5	4,248.1	8,500.0	34,448.4	1,000,000.0	22,122,179
Internal credit score	5.42	2.85	1	1	3	5	8	10	10	22,122,179
Male indicator	0.70	0.46	0	0	0	1	1	1	1	22,122,179
Age	27.43	6.36	18	19	23	26	31	46	60	22,122,179
Bachelor’s degree or more indicator	0.13	0.34	0	0	0	0	0	1	1	22,122,179

Panel B. RD subsample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	# Obs.
Overdue amount (yuan)	142.13	112.53	20.01	22.12	58.60	106.09	188.68	503.16	848.18	1,010,540
Remaining principal (yuan)	304.74	112.45	100.00	104.03	209.00	308.03	400.66	496.19	499.99	1,010,540
Internal credit score	4.91	2.77	1	1	3	4	7	10	10	1,010,540
Male indicator	0.72	0.45	0	0	0	1	1	1	1	1,010,540
Age	26.81	5.98	18	19	22	25	30	46	60	1,010,540
Bachelor’s degree or more indicator	0.10	0.31	0	0	0	0	0	1	1	1,010,540

Panel C. Completely randomized subsample

Variable	Mean	S.D.	Min	P1	P25	P50	P75	P99	Max	# Obs.
Overdue amount (yuan)	1,522.9	1,846.4	20.2	86.3	554.8	1,018.0	1,849.4	8,653.9	35,639.9	147,424
Remaining principal (yuan)	8,593.9	6,966.4	300.1	467.5	3,438.0	6,600.1	11,667.8	30,968.8	34,919.6	147,424
Internal credit score	5.97	2.71	1	1	4	6	8	10	10	147,424
Male indicator	0.70	0.46	0	0	0	1	1	1	1	147,424
Age	27.77	6.79	18	19	22	26	32	47	59	147,424
Bachelor’s degree or more indicator	0.10	0.31	0	0	0	0	0	1	1	147,424

Table 3. Treatment effect of permanent assignment to AI callers – RD results.

Panel A reports tests of continuity at the 300-yuan remaining principal threshold for predetermined loan and borrower characteristics measured on day 2 past due. Columns 2 and 3 report the regression-fitted value of the variables of interest at the threshold from the left side (permanent AI) and the right side (human). Column 4 reports the differences between the left and right fitted values (AI minus human), with z-statistics, p-values, and 95% robust bias-corrected RD confidence intervals in the following columns. Panel B reports the treatment effect of being permanently assigned to an AI caller on the present value of cash flows collected from the case within a given horizon after the original due date, scaled by the day 2 overdue balance. In addition to the first seven columns as in Panel A, Panel B column 8 re-estimates the differences around the threshold while controlling for all five covariates (overdue payment in logs and credit scores using decile dummies) in Panel A. Local linear regressions with uniform kernels are used to obtain point estimates in all rows, and robust bias-corrected RD z-statistics are estimated by local quadratic regressions. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)		(8)
	Variable	Left Mean (AI)	Right Mean (Human)	Diff. (L–R)	z-stat.	p-val.	95% Robust Bias-Corrected C.I.		Diff. with Covariates
Panel A. Loan and borrower characteristics									
(1)	Overdue payment	142.9	143.8	-0.94	-0.40	0.686	-6.83	4.50	
(2)	Credit score	4.96	4.90	0.06	1.30	0.194	-0.03	0.15	
(3)	Male	0.714	0.722	-0.008	-1.51	0.132	-0.018	0.002	
(4)	Age	26.63	26.65	-0.027	-0.22	0.824	-0.199	0.159	
(5)	Bachelor's degree or higher	0.105	0.108	-0.003	-0.73	0.468	-0.011	0.005	
Panel B. Cumulative NPV collected									
(6)	2 days	0.168	0.206	-0.038***	-9.02	<0.001	-0.044	-0.028	-0.040***
(7)	5 days	0.418	0.468	-0.050***	-7.77	<0.001	-0.059	-0.035	-0.049***
(8)	10 days	0.581	0.658	-0.078***	-12.73	<0.001	-0.087	-0.064	-0.076***
(9)	30 days	0.733	0.824	-0.092***	-18.40	<0.001	-0.099	-0.080	-0.091***
(10)	60 days	0.782	0.857	-0.075***	-17.24	<0.001	-0.082	-0.065	-0.074***
(11)	90 days	0.800	0.864	-0.065***	-16.06	<0.001	-0.071	-0.056	-0.064***
(12)	180 days	0.820	0.872	-0.052***	-13.84	<0.001	-0.058	-0.044	-0.052***
(13)	360 days	0.828	0.875	-0.046***	-12.59	<0.001	-0.052	-0.038	-0.046***

Table 4. Treatment effect of five days of AI callers – Completely randomized subsample.

This table compares two groups in the 10% completely randomized subsample: the treatment group handled by AI callers from days 2 to 5 past due before being given to human callers on day 6, and the control group handled by human callers from day 2 onwards. Panel A reports tests that the two groups are balanced in their pre-determined loan and borrower characteristics measured on day 2. Columns 2 and 3 report the average of the variables of interest among cases assigned to the treatment and the control groups, respectively. Column 4 reports the differences between the averages, estimated by regressions that also control for day-2 calendar month fixed effects, with *t*-statistics in column 5. Panel B reports the performance of the two groups measured by collected NPV scaled by the day 2 overdue balance. The treatment effect estimates come from regressions of the variable of interest onto a treatment group indicator and day-2 calendar month dummies. In columns 2-5, cases are equally weighted in the regression, and in column 6, cases are weighted by their overdue payment amount on day 2. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1) Variables	(2) Mean (AI + Human)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.	(6) Diff: AI – Human (value-wgt.)
Panel A. Loan and borrower characteristics						
(1)	Overdue amount	1524.2	1521.6	1.5	0.15	
(2)	Remaining principal	8585.2	8602.6	-19.5	-0.54	
(3)	Internal credit score	5.971	5.960	0.009	0.66	
(4)	Male	0.701	0.703	-0.002	-0.83	
(5)	Age	27.75	27.79	-0.043	-1.25	
(6)	Bachelor's degree or higher	0.104	0.104	-0.001	-0.36	
Panel B. Cumulative NPV collected						
(7)	2 days	0.180	0.268	-0.088***	-42.89	-0.110***
(8)	5 days	0.389	0.510	-0.121***	-50.07	-0.144***
(9)	10 days	0.612	0.635	-0.023***	-10.02	-0.023***
(10)	30 days	0.734	0.741	-0.0076***	-3.77	-0.0074***
(11)	60 days	0.768	0.774	-0.0069***	-3.63	-0.0079***
(12)	90 days	0.783	0.790	-0.0067***	-3.69	-0.0067***
(13)	180 days	0.796	0.803	-0.0068***	-3.86	-0.0068***
(14)	360 days	0.802	0.809	-0.0069***	-4.02	-0.0069***

Table 5. Phone call outcomes of AI and human callers on day 2.

This table compares outcomes for phone calls made on day 2 past due by AI callers versus human callers within the completely randomized subsample. Columns 2 and 3 report the average of the variables of interest within the AI and human caller groups, respectively. Column 4 reports the differences between the averages, with *t*-statistics in the following column. The estimates of the differences are based on regressions of the outcome variable onto an AI-caller indicator and calendar month dummies. Panel A uses data on all phone calls made on day 2 past due, while Panels B and C restrict the sample to the first call answered by each borrower on day 2. The timing adjustment accounts for the time of the call by additionally controlling for fixed effects for the hour of the call. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Panel A. All calls on day 2 past due.

	(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
(1)	# Phone calls per borrower	4.75	6.12	-1.38***	-87.59
(2)	# Phone calls answered	0.65	1.00	-0.35***	-57.63
(3)	% Phone calls answered	0.137	0.163	-0.026***	-19.31

Panel B. First answered calls.

	(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human	(5) *t*-stat.
(1)	Time of calls	11:47 A.M.	11:31 A.M.	17 minutes***	12.51
(2)	Ringing time to answer (sec)				
	Unadjusted	19.50	20.69	-1.25**	-11.66
	Timing-adjusted			-0.02	-0.17
(3)	Duration (sec)				
	Unadjusted	28.02	47.22	-19.02***	-41.41
	Timing-adjusted			-30.96***	-61.25
(4)	% Promise to repay				
	Unadjusted	0.442	0.650	-0.211***	-52.44
	Timing-adjusted			-0.211***	-45.86

Panel C. Repayment after first answered calls.

(1) Variables	(2) Mean (AI)	(3) Mean (Human)	(4) Diff: AI – Human (Timing-adjusted)	(5) *t*-stat. (Timing-adjusted)
Repay (fully or partially) after the first answered call within …				
15 minutes	0.048	0.045	-0.003	-1.32
30 minutes	0.064	0.072	-0.017***	-6.94
1 hour	0.084	0.115	-0.042***	-14.48
2 hours	0.114	0.181	-0.081***	-23.72
5 hours	0.174	0.305	-0.141***	-34.63
the same day	0.273	0.448	-0.181***	-40.03

Table 6. Repayment after the first answered calls, conditional on promises to repay.

This table reports the fraction of borrowers who repay their debts (fully or partially) within various periods after the first answered phone call from AI or human callers on day 2 after the due date, conditioning on whether the borrowers make a promise to repay their debt during the conversation. The analysis uses the completely randomized subsample. Columns (a) to (d) reports simple group averages and columns (e) to (f) reports the differences between them. The estimates of the differences are based on regressions of the outcome variable onto an AI-caller indicator, a promise-to-repay indicator, the interaction of the two indicators, and calendar month and hour-of-call dummies. The numbers in parentheses are *t*-statistics for the difference between the proportions in the indicated columns. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	With a Promise		Without a Promise				
	(a) AI	(b) Human	(c) AI	(d) Human	(e) Diff: (a) – (c)	(f) Diff: (b) – (d)	(g) Diff: (e) – (f)
15 minutes	0.063	0.056	0.036	0.025	0.029***	0.032***	-0.003
					(11.05)	(12.07)	(-0.92)
30 minutes	0.085	0.092	0.047	0.037	0.039***	0.055***	-0.016***
					(12.71)	(17.62)	(-3.74)
1 hour	0.112	0.144	0.061	0.060	0.052***	0.084***	-0.032***
					(14.46)	(22.83)	(-6.25)
2 hours	0.155	0.226	0.080	0.098	0.076***	0.127***	-0.050***
					(18.02)	(29.17)	(-8.3)
5 hours	0.234	0.377	0.126	0.171	0.110***	0.201***	-0.091***
					(21.92)	(39.15)	(-12.72)
the same day	0.344	0.539	0.217	0.279	0.132***	0.247***	-0.115***
					(23.67)	(43.41)	(-14.55)

Table 7. Relationship between phone call strategies and performance on day 2.

This table shows how human callers' phone call frequency is related to collection performance on day 2 past the due date. For each frequency metric, human callers specializing in the first five days past due are grouped into terciles by their average call frequency measure (indicated in the column supertitles) across all their assigned cases in the second or later delinquencies in month $t-1$. The column "Avg. in Month *t*" contains the average day 2 call frequency (from any caller) in month t to borrowers assigned at the beginning of day 2 to a caller in the given tercile. AI callers' average call frequency is reported in the last row of each panel. The columns "Avg. Diff." and "NPV2 Diff." contain the difference in average call frequency or collected scaled payments between the medium or high tercile and the lowest tercile. The borrower sample is the completely randomized subsample. Panel A considers all phone calls and all borrowers, while Panel B restricts the sample to borrowers who make a promise to repay on day 2. Numbers in parentheses are *t*-statistics. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Panel A. Phone call frequency.

	# Phone Calls (Per Borrower-Day)			Average Time Interval Between Calls (Minutes)		
Human Tercile in Month $t-1$	Avg. in Month t	Avg. Diff. (t-stat.)	NPV2 Diff. (t-stat.)	Avg. in Month t	Avg. Diff. (t-stat.)	NPV2 Diff. (t-stat.)
1 (Low)	6.62			94.35		
2 (Med)	6.82	0.20***	-0.014**	99.18	4.83***	-0.010*
		(4.41)	(-2.38)		(4.99)	(-1.80)
3 (High)	6.98	0.36***	0.006	105.39	11.04***	-0.014**
		(7.97)	(1.05)		(11.28)	(-2.41)
AI	4.75			112.08		

Panel B. Follow-up frequency within the same day, conditional on a promise to repay on day 2.

	% Follow-Up Within the Same Day			Average Time Interval to Follow-Up After Receiving a Promise (Minutes)		
Human Tercile in Month $t-1$	Avg. in Month t	Avg. Diff. (t-stat.)	NPV2 Diff. (t-stat.)	Avg. in Month t	Avg. Diff. (t-stat.)	NPV2 Diff. (t-stat.)
1 (Low)	0.862			88.38		
2 (Med)	0.878	0.015*	-0.002	94.61	6.23**	-0.004
		(1.73)	(-0.29)		(2.24)	(-0.71)
3 (High)	0.881	0.019**	0.005	107.16	18.77***	-0.006
		(2.09)	(0.79)		(6.69)	(-1.03)
AI	0.829			214.39		

Table 8. Probability of delinquency on subsequent monthly payments.

This table shows borrowers' probability of missing subsequent monthly payments after they are initially assigned to the treatment group (five days of AI followed by humans) or the control group (always human) in the completely randomized subsample. Subsequent monthly payments due are identified by the number of months between the payment in question and the month borrowers entering the completely randomized subsample. Columns 2 and 3 report the fractions of borrowers who miss the *p*th payment in the treatment and the control groups, respectively. Column 4 reports the difference (AI minus human) and column 5 the corresponding *t*-statistic based on regressions of the outcome variable onto an AI-caller indicator and calendar month dummies. The sample for the *p*th payment only includes loans with at least *p* months of payments between the month of entry into the completely randomized subsample and the loan's maturity. The number of such loans are reported in the last column. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

(1) Subsequent Monthly Payment	(2) AI + Human	(3) Human	(4) Diff.: AI – Human	(5) *t*-stat.	(6) # Loans
1	0.627	0.604	0.023***	8.99	143,481
2	0.602	0.579	0.022***	8.48	139,211
3	0.589	0.570	0.019***	7.23	133,978
4	0.575	0.556	0.019***	7.02	129,020
5	0.562	0.545	0.017***	6.15	122,278
6	0.557	0.539	0.019***	6.42	116,722
7	0.545	0.529	0.016***	5.44	111,997
8	0.534	0.517	0.018***	5.81	105,727
9	0.527	0.511	0.016***	5.11	98,267
10	0.513	0.499	0.014***	4.17	89,937
11	0.504	0.494	0.011***	2.99	80,073
12	0.505	0.492	0.013***	3.64	72,828

Table 9. Day of the week of first call's effect on cumulative NPV collected.

This table compares borrowers always called by humans who are first called on weekends versus weekdays in the completely randomized subsample. Columns 1 and 2 contain average outcomes of borrowers first contacted on weekends and weekdays, respectively. Column 3 is the difference between the two types of borrowers, and column 4 reports the corresponding *t*-statistic, based on regressions of the variable of interest onto a weekend indicator and week fixed effects. Standard errors are clustered at the calendar month level. Columns 5 and 6 re-estimate the outcome differences additionally controlling for log overdue payment amount, internal credit score decile dummies, gender, age, and attainment of a bachelor's degree measured on day 2 after the due date. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	(1) Weekend Average	(2) Weekday Average	(3) Diff: (1) – (2)	(4) *t*-stat.	(5) Diff: (1) – (2) w/ Covar.	(6) *t*-stat.
Call outcomes on day 2:						
% Calls answered	0.161	0.164	-0.004	-1.17	-0.003	-0.94
# Calls per borrower	6.204	6.087	0.106*	1.93	0.110**	2.29
# Calls answered	0.999	1.001	-0.005	-0.25	0.000	0.00
Call duration (sec)	73.06	71.78	0.40	0.22	1.00	0.60
% Promise to repay	0.666	0.666	0.003	0.38	0.005	0.56
Cumulative NPV collected:						
2 days	0.258	0.272	-0.015***	-3.86	-0.014***	-3.92
3 days	0.376	0.391	-0.017***	-3.86	-0.016***	-4.24
4 days	0.454	0.464	-0.011**	-2.64	-0.011**	-2.71
5 days	0.508	0.511	-0.005	-1.09	-0.004	-0.98
6 days	0.556	0.555	0.000	0.10	0.001	0.26
10 days	0.633	0.635	-0.002	-0.46	-0.001	-0.38
15 days	0.681	0.681	-0.000	-0.03	0.000	0.03
30 days	0.741	0.741	-0.001	-0.26	-0.001	-0.30
60 days	0.774	0.775	-0.001	-0.38	-0.001	-0.51
90 days	0.790	0.789	0.000	0.04	-0.000	-0.07
180 days	0.803	0.803	-0.001	-0.15	-0.001	-0.27
360 days	0.808	0.809	-0.001	-0.27	-0.001	-0.39

Table 10. Caller work experience's effect on cumulative NPV collected.

This table compares borrowers assigned to senior versus junior callers at the beginning of day 2. The sample is the completely randomized subsample restricted to borrowers initially assigned to human callers specializing in debts in the first 5 days past due. In Panel A, Columns 1 and 2 contain the cumulative NPV collected from debts first assigned to the two types of callers. Column 3 shows the difference between the two, and column 4 contains the corresponding t-statistics. The estimates of the differences are based on regressions of the variable of interest onto a senior-caller indicator and calendar month dummies, clustering standard errors at the month level. Columns 5 and 6 re-estimate the outcome differences additionally controlling for log overdue payment amount, internal credit score decile dummies, gender, age, and attainment of a bachelor's degree measured on day 2. Panel B further splits junior callers into two groups according to their work experience level and reports differences relative to senior callers that control for the additional covariates. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Panel A. Debt collection outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Junior Avg.	Senior Avg.	Diff: (1) – (2)	t-stat.	Diff: (1) – (2) w/ Covar.	t-stat.
Cumulative NPV collected:						
2 days	0.239	0.253	-0.009	-1.66	-0.011**	-2.35
3 days	0.365	0.365	-0.005	-0.74	-0.008	-1.25
4 days	0.443	0.437	-0.002	-0.34	-0.005	-0.93
5 days	0.495	0.486	0.002	0.39	-0.001	-0.20
6 days	0.543	0.532	0.002	0.43	-0.001	-0.15
10 days	0.628	0.613	0.003	0.64	0.001	0.19
30 days	0.738	0.719	-0.002	-0.33	-0.003	-0.56
60 days	0.773	0.752	0.001	0.20	0.000	0.02
90 days	0.788	0.767	-0.002	-0.52	-0.003	-0.70
180 days	0.802	0.780	-0.002	-0.55	-0.003	-0.69
360 days	0.808	0.786	-0.002	-0.58	-0.003	-0.71

Panel B. Splitting the group of junior callers.

	(1)	(2)	(3)	(4)	(5)
Work Experience:		0 to 1 month		2 to 3 months	
	Senior Avg.	Diff.	t-stat.	Diff.	t-stat.
Cumulative NPV collected:					
2 days	0.253	-0.015***	-2.91	-0.009	-1.59
3 days	0.365	-0.011*	-1.75	-0.006	-0.82
4 days	0.437	-0.013**	-2.17	-0.001	-0.15
5 days	0.486	-0.006	-0.92	0.002	0.28
6 days	0.532	-0.005	-0.68	0.001	0.20
10 days	0.613	-0.002	-0.36	0.002	0.42
30 days	0.719	-0.007	-1.24	-0.001	-0.18
60 days	0.752	-0.005	-0.94	0.003	0.71
90 days	0.767	-0.008	-1.59	-0.000	-0.08
180 days	0.780	-0.006	-1.43	-0.001	-0.25
360 days	0.786	-0.006	-1.64	-0.001	-0.22

Table 11. Performance of different versions of AI callers.

This table reports the cumulative collected NPV differences between consecutive versions of AI callers at the horizons of 2-6, 8, and 10 days past due. The differences are estimated by linear regressions of collected NPV on indicators for which AI version the case was assigned to and calendar month fixed effects. The last row reports the average cumulative NPV collected at each horizon for all AI-assigned cases in the completely randomized subsample. Numbers in parentheses are *t*-statistics. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

	NPV Horizon (days past due)						
Version Diff.	2	3	4	5	6	8	10
V2 – V1	-0.0052	0.0084	0.0063	0.0145	0.0064	0.0040	0.0071
	(-0.53)	(0.68)	(0.51)	(1.17)	(0.49)	(0.32)	(0.57)
V3 – V1	0.0127*	-0.0041	-0.0060	-0.0018	-0.0109	-0.0119	-0.0144
	(1.74)	(-0.46)	(-0.65)	(-0.2)	(-1.12)	(-1.29)	(-1.58)
V4 – V1	0.023***	0.0242**	0.0234**	0.0303***	0.0118	0.005	-0.0013
	(2.67)	(2.34)	(2.17)	(2.71)	(1.05)	(0.46)	(-0.12)
V5 – V4	0.0218*	0.0133	0.0072	-0.0047	0.0045	0.01	0.0052
	(1.92)	(1.04)	(0.49)	(-0.33)	(0.33)	(0.74)	(0.37)
V6 – V5	-0.001	-0.0034	0.0062	0.0043	0.0119	0.0144	0.0163*
	(-0.14)	(-0.38)	(0.75)	(0.45)	(1.31)	(1.56)	(1.91)
Average NPV	0.180	0.279	0.342	0.389	0.490	0.570	0.612