$1 + 1 > 2$? Information, Humans, and Machines

Tian Lu
Arizona State University, lutian@asu.edu

Yingjie Zhang
Peking University, yingjiezhang@gsm.pku.edu.cn

With the explosive growth of data and the rapid rise of artificial intelligence (AI) and automated working processes, humans inevitably fall into increasingly close collaboration with machines, either as employees or consumers. Problems in human-machine interaction arise as a consequence, not to mention the dilemmas posed by the need to manage information on ever-expanding scales. Considering the general superiority of machines in this latter respect, compared to human performance, it is essential to explore whether human–machine collaboration is valuable, and if so, why. Recent studies have proposed diverse explanation methods to uncover machine learning algorithms' "black boxes," aiming to reduce human resistance and enhance efficiency. However, the findings of this literature stream have been inconclusive. Little is known about the influential factors involved or the rationale behind their impacts on human decision processes. We aimed to tackle the above issues in the present study by specifically examining the joint impact of information complexity and machine explanations. Specifically, we cooperated with a large Asian microloan company to conduct a two-stage field experiment. Drawing upon studies in dual-process theories of reasoning that have proposed different conditions necessary to arouse humans' active information processing and systematic thinking, we tailored the treatments to vary the level of information complexity, the presence of collaboration, and the availability of machine explanations. We observed that with large volumes of information and with machine explanations alone, human evaluators could not add extra value to the final collaborative outcomes. However, when extensive information was coupled with machine explanations, human involvement significantly reduced the default rate compared with machine-only decisions. We disentangled the underlying mechanisms with three-step empirical analyses. We revealed that the co-existence of large-scale information and machine explanations can invoke humans' active rethinking, which in turn, shrinks gender gaps and increases prediction accuracy. In particular, we demonstrated that humans could spontaneously associate newly emerging features with others that had been overlooked but had the potential to correct the machine's mistakes. This capacity not only underscores the necessity of human-machine collaboration but also offers insights into system designs. Our experiments and empirical findings provide non-trivial implications that are both theoretical and practical.

Key words: Decision-making, Gender Biases, Human-Machine Collaboration, Information Processing, Machine Explanations, Micro-finance, Rethinking

1. Introduction

Given the fast rate of artificial intelligence (AI) commercialization and its penetration into daily life, humans have started to closely collaborate with machines, both as employees and consumers (Alibaba 2018, Wang et al. 2023a). For example, many companies have introduced AI-based coaching systems to assist humans and improve their decision-making effectiveness and efficiency (Loutfi 2019). In reality, humans and machines can complement each other. Previous research has found that the decision-making accuracy of machine-learning algorithms is generally higher than that of humans under normal circumstances (Grove et al. 2000). However, humans are more likely to use experience to identify and process low-frequency cases that are difficult to include in machine-learning algorithms; humans also have more advantages than machines in terms of flexibility (Sawyer 1966). More importantly, humans' deep thinking is a well-established and well-understood tool for augmenting performance on independent or team tasks (Amit and Sagiv 2013).

Unfortunately, there are various constraints, such as information opacity, machine-learning algorithms' complexity, and personnel's lack of experience with or understanding of advanced technologies. Accordingly, the realized performance of human-machine collaborations falls short of the expectation due to distrust of machines (Jacovi et al. 2021) or over-reliance on them (Fügener et al. 2021). Even worse, without properly designed collaboration systems, humans' involvement could reduce the collaborative performance for various reasons, such as their being over-cautious (Lu et al. 2023b) or hyper-focused on details (Wang et al. 2023c).

To address the urgent, essential question regarding how to efficiently change humans' responses to machines from either aversion or over-reliance to active contribution, researchers have recently begun to turn to machine-learning model explanations (Schmidt et al. 2020, Bauer et al. 2023). However, previous investigations in this vein have predominantly concentrated on technical solutions and lacked a comprehensive examination of the conditions and underlying mechanisms that influence the solutions' impact on human decision processes. This omission introduces certain limitations, as not all model explanations prove effective in every scenario (Chen et al. 2023).

In this study, emphasis is placed on task complexity, particularly information complexity, a contingent factor that plays a pivotal role in shaping the effectiveness of machine explanation implementations. We posit that task complexity and machine explanations should work concurrently to foster deep thinking in humans, thereby contributing to the efficacy of human-machine collaborations. Specifically, task complexity and information richness engage humans in deliberate information processing by capturing their attention and interest in complex decision tasks (Levin et al. 2000). The presentation of machine explanations that serve as valuable cues and decision-making references prompt humans to carefully reassess decisions, address conflicts, and actively process information through cognitive reasoning (Mantel and Kardes 1999). Through the alignment of these conditions, humans are more likely to employ enhanced decision-making strategies, ultimately improving the performance of human-machine collaborations.

Notably, prior studies exploring the value of machine explanations have typically conducted lab experiments or simulations alone. This approach proves challenging, as participants tend to present differently and participate more actively in a controlled lab environment (Keil et al. 2000). Consequently, there is a compelling need to adopt a more pragmatic approach−−a realization that led us to design and implement field experiments. These experiments serve as a crucial means of observing and analyzing human behavior in more authentic, real-world scenarios, particularly with regard to their ability to navigate and respond to varying levels of information complexity and cues.

Therefore, in this paper, we apply field experiments to determine whether and how humans' potential to achieve "$1 + 1 > 2$" can be realized, particularly in the context of increasing technological development and human-machine collaboration. Our three research questions are: (1) What is the realized performance when humans and machines collaborate under different levels of information complexity and different system designs? (2) What are the underlying mechanisms? (3) How do human characteristics affect collaborative performance?

We focused on the microloan industry and partnered with a large Asian microloan company to conduct a two-stage field experiment. We dove into the dual-process theories of reasoning (Evans 2003), suggesting two prerequisites for invoking humans' deep thinking: *information complexity* initially draws humans' attention and engages them in the tasks, and *useful cues* drive humans to actively consider the task. Accordingly, we experimentally manipulated how much information about borrowers was provided to evaluators, whether evaluators got to see the machine's recommendation, and whether the machine's recommendation was explained to the evaluators.

Our empirical analyses yielded several interesting findings. *First*, with small information volumes, human evaluators could not add extra value to the final outcome (i.e., the default rate prediction accuracy). *Second*, the human evaluators outperformed the machines when the human evaluators were allowed to observe the machine's suggestions before making their final decisions and when the machine explanations were offered and the information volume was large. In these cases, human evaluation resulted in a 2.02% reduction in the default rate (from 5.15 to 3.13%). However, this improvement disappeared if either machine explanations or information complexity were not given. *Third*, we observed that when humans and machines made decisions independently, a certain amount of disagreement was inevitable. In the human-machine collaboration modes, a disagreement of 62.82% resulted from a small information volume without machine explanations, compared with 85.67% disagreement resulting from large amounts of information and disclosure of machine explanations.

To disentangle the potential mechanisms and explain the above findings, we employed a three-step analytical framework. Our findings suggested several important insights. *First*, human evaluators tended to stick with traditionally important features such as income or education level, while machines explored more possibilities using other sources of information, including shopping and offline trajectory behavior. This explains why machines, in general, performed better than humans, especially when large amounts of information were offered. *Second*, with the availability of machine explanations and large information volumes, evaluators performed active rethinking when inconsistent decisions were made. This improved their final decision accuracy by, for example, correcting the risk evaluation of female borrowers. However, such a rethinking process did not occur if either condition was not satisfied. *Third*, we disentangled the "rethinking" procedure in which humans associate the machine explanations with other features if they considered the displayed features to be "non-informative".

Furthermore, when considering individual heterogeneity among human evaluators, we found that though more experienced evaluators were less likely to follow the machines' suggestions, they were stimulated in their rethinking by the machines' suggestions and explanations, and this, in turn, improved company performance. In addition, we compared repayment behavior to examine the existence of potential gender-based decision biases. Our findings suggest that with more data and machine explanations, human-machine collaboration could potentially shrink the inter-gender default rate gap, which was initially and unintentionally produced by machine-learning algorithms. This further highlights the value and necessity of collaboration between humans and machines.

The contributions of our study are multi-fold. *First*, it adds to the emerging literature on human–machine collaboration. Whereas a few of the most recent studies have investigated whether humans and machines complement each other in decision-making in different contexts (e.g., Cao et al. 2021, Luo et al. 2019, Zhang et al. 2023), the majority have suggested outcomes only implicitly or ostensibly. Through in-depth mechanism detection analyses, our study unravels how and why properly designed collaboration can invoke humans to contribute. Thus, we advance this stream of literature by revealing the existence and value of humans' rethinking processes, both theoretically and empirically. *Second*, we contribute to the recent literature on the value of offering machine explanations within the context of human-machine collaboration. The existing literature has not reached a consensus on how humans respond to machines' advice in the case of machine explanations (Krishna et al. 2022). Our study proposes and verifies one reason of inconclusive findings in prior literature: the outcome of providing machine explanations is related to other conditions such as humans' perception of the environmental or task complexity. Whereas previous studies have largely suggested that displaying (feature-based) machine explanations would invoke humans' System 1 thinking (i.e., heuristics or rules-of-thumb for making quick judgments) rather than System 2 (active reasoning and rethinking) (Chen et al. 2023), we demonstrate that with a proper collaboration design, machine explanations can prompt humans' rethinking and improve human–machine collaboration. *Third*, we add to the recent stream of literature regarding machine biases. Recent studies have proposed the utilization of multi-source data to alleviate algorithmic discrimination and sample biases. In fact, there is evidence that alternative data sources would eliminate biases related to race and socioeconomic factors (Lu et al. 2023a). However, machine failure has already been proven (Fuster et al. 2022, Hu et al. 2022), so this paper not only identifies the sources of gender biases but also uncovers the value and necessity of human involvement to make up for machine failure.

2. Related Studies

This section first summarizes three related streams of literature, then offers an introduction to the theoretical framework underpinning experimental treatment design.

2.1. Human Collaboration With and Aversion to Machines

AI applications require human intervention and assistance. Previous studies have explored the pros and cons of human–machine collaboration in decision-making. For example, studies have shown that most statistical models exceed or approach the judgment accuracy of the average clinician (Camerer et al. 2019). Machine algorithms have been extensively shown to manage substantial amounts of data more proficiently than humans (Peukert et al. 2023, Wang et al. 2023c). However, despite the fact that machines can make highly accurate predictions, it is difficult for them to handle random or uncertain cases and boundary cases whose features show contradictory patterns on the prediction objectives (labels) (Guo and Wang 2015). By contrast, humans are found to be better at identifying rare cases (Sawyer 1966) and to perform more effectively in innovative areas such as new product development (Lou and Wu 2021). Recent studies have shown the superiority of human–machine collaborations over both full machine automation and human-only operations (Fügener et al. 2022), and have shed light on the merits of "the human-in-the-loop" (Fügener et al. 2021). On the one hand, machines can augment the capabilities of humans, such as managers (Davenport et al. 2020); and on the other hand, humans can complement machines by contributing their general intelligence (Te'eni et al. 2023) and diverse ideas (Wang et al. 2023d, Zhang et al. 2023) and incorporating private information (i.e., data that only humans can use such as in-house data) (Choudhury et al. 2020, Ibrahim et al. 2021, Sun et al. 2022). Cao et al. (2021) showed that when analysts are given access to a small amount of alternative data and in-house machine resources, combining machines' computational power and humans' understanding of soft information produces the best performance in generating accurate forecasts.

However, recent research has also revealed that humans might resist the adoption or usage of machines, resulting in low efficiency of human–machine collaboration (Allen and Choudhury 2022, de Véricourt and Gurkan 2023, Wang et al. 2023b). This resistance exists not only among those who accept machines' advice (e.g., Commerford et al. 2022, Liu et al. 2023), but also among machine-based service targets, namely ordinary consumers. For example, the adoption of chatbots has had negative effects on user acceptance and efficiency due to consumers' insufficient knowledge and relative lack of empathy from chatbots (Luo et al. 2019). However, this negative impact may be mitigated by users' experience levels (Luo et al. 2021, Tong et al. 2021), flexibility, and willingness to make adjustments based on machines' predictions (Dietvorst et al. 2018). Human aversion to machines could also be due to the potential of machines to threaten human jobs. AI robots have replaced and will replace human labor in different ways in various fields (Brynjolfsson and Mitchell 2017, Lu et al. 2018). Machines have outperformed humans in many jobs, especially low-skilled, repetitive, and dangerous ones (Autor and Dorn 2013). Conversely, Fügener et al. (2021) warned that we must also attend to humans' over-reliance on machines, which would render human–machine collaboration useless.

2.2. Machine Explanations

The lack of model explanations could result in human aversion to machines, stemming from a sense of distrust (Siau and Wang 2018). To avoid such negative outcomes, the existing literature has examined multiple approaches. A commonly adopted approach improves trust in human–machine collaboration settings by offering more detailed information of machine-learning decisions (Lu et al. 2019, Rai 2020). Through various post-hoc explanation methods, human participants can be assisted in constructing suitable mental models under diverse conditions, thereby enhancing their trust and the model efficiency (Mohseni et al. 2020). However, this approach should be employed with caution. Schmidt et al. (2020) indicated that offering unintuitive explanations (i.e., those dealing with features humans are unfamiliar with) may fail to boost humans' trust in machines. Rudin (2019) also cautioned that post-hoc explanations tend to offer incomplete and biased information regarding the mechanisms underlying algorithms. This may lead participants to overestimate their ability to explain decisions declaratively, resulting in misinformation.

Our research aligns with this common practice. However, while some previous studies have explored the impact of machine explanations on human–machine collaboration, few have delved into the specific mechanisms of how and why such an approach works in influencing human decision processes. The most similar study to ours is Bauer et al. (2023), which revealed that humans can dynamically adjust the importance they attribute to available information and adapt their mental models based on machine explanations. Additionally, their findings highlighted that the provision of machine explanations might reinforce confirmation bias, potentially resulting in suboptimal or biased decisions. However, our study differs from Bauer et al. (2023) in at least two key aspects. *First*, while Bauer et al. (2023) only attended to a limited number of borrower features, we additionally consider information complexity. As outlined in Section 2.4, we contend that the effectiveness of machine explanations in shaping individuals' information processing depends on the complexity of the information presented to them. Machine explanations stimulate active cognitive information processing only under specific conditions of information complexity. Furthermore, under certain conditions, the overall performance of human–machine collaboration may see improvement rather than deterioration. *Second*, the findings of the study by Bauer et al. (2023) could have been influenced by their use of online lab experiments. By their nature, lab experiments present challenges related to sample representativeness (Compeau et al. 2012). Of greater significance is the potential for participants to react differently within the confines of a lab setting, which is characterized by specific monitoring and anchoring conditions. Participants might naturally respond more actively and attentively to the experimental manipulations, potentially leading to an overestimation of their behavioral outcomes (Keil et al. 2000). In contrast, our study adopts a field experiment approach within a real-world micro-finance context to examine individuals' decision-making in a more natural setting.

2.3. Investors' Decision-Making in Micro-finance

Many scholars have focused on individual investors' decision-making in micro-finance businesses, including P2P lending, crowdfunding, and microloans. A subset of the literature has revealed the important factors that investors consider in their decision-making (e.g., Gonzalez and Loureiro 2014, Tao et al. 2017, Wang et al. 2019). Studies have also identified biases in micro-finance investors' decisions, including preferences regarding gender (Chen et al. 2017) or location (Lin and Viswanathan 2016). Recent research has paid attention to the value of machine-assisted tools in financial decision-making. For example, Ge et al. (2021) found that P2P lending investors experiencing more defaulted loans are more likely to perceive the market to be risky and thus tend to rely more on their own judgment rather than a robot advisor. Additionally, some investors attempt to intervene in machine usage. They may be more concerned about returns and less likely to lose confidence in machines immediately after observing a machine failure (Germann and Merkle 2019), or they may tend to adjust their machine usage based on the latest performance (Ge et al. 2021). In our study, we also delve into both decision-making accuracy and potential biases within the micro-finance context. However, unlike existing studies, our emphasis lies in examining how machine decisions function as recommendations to influence users' decision-making.

2.4. Theoretical Underpinning: The Dual-Process Theories of Reasoning

Humans' and machines' respective advantages in decision-making and their collaborative value lie in their complementarity (Feuerriegel et al. 2022). However, humans fall easily into aversion toward or over-reliance on machines; neither situation yields better decision outcomes than either human-only or machine-only decision-making. Therefore, one key to promoting the value of collaboration between humans and machines is to invoke humans' deep thinking in their co-working with machines. The literature on dual-process theories of reasoning (Evans 2003), our theoretical underpinning, raises the question of how humans' deep thinking can be aroused in machine-assisted tasks. The dual-process theories of reasoning propose the existence of two cognitive systems, "System 1" and "System 2", that underlie thinking and reasoning. System 1 processes information and reasoning fast, automatically, and with minimal effort, leading to quick and instinctive decision-making as a rapid response to familiar situations and stimuli. In contrast, System 2 operates at a slower pace, involves deliberate thought, and requires conscious effort. It incorporates logical reasoning and analysis and involves the application of cognitive resources (Kahneman 2011).

Several factors can determine whether individuals opt for System 1 or System 2 information processing and reasoning. To encourage individuals to embrace System 2 processing, certain conditions must be met. Specifically, since System 2 is typically involved in complex tasks, problem-solving, critical thinking, and decision-making in novel or challenging situations, task complexity is a primary condition. Task complexity, often represented by information complexity (Amit and Sagiv 2013), stimulates deep thinking in individuals by capturing their attention and interest in decision tasks (Levin et al. 2000). As proposed by Endsley (1995), being well-informed about the situation at hand is a prerequisite for subsequent deep reasoning and action selection. Information complexity, manifested as multiple alternatives and/or numerous attributes, influences users' situational processing of observed information (Bauer et al. 2023, Sun and Taylor 2020). Specifically, new attributes provide novel pieces of information that enhance one's recognition of the decision tasks and domain (He et al. 2020). Faced with greater volumes of more diverse, unfamiliar information, individuals are inclined to invest more effort in reasoning through more ambiguous task situations (Van der Schalk et al. 2010). In other words, although more complex information may not necessarily result in increased decision-making accuracy, it does enhance individual's willingness to actively participate in decisions (Oskamp 1965). With complex information, people are more willing to perceive the increase in information as useful and desirable, even if it comes with a certain level of burden (Amit and Sagiv 2013). In contrast, with simple information, people tend to make rapid decisions via System 1 processing (Speier 2006).

The second condition for motivating people to engage in high-quality System 2 processing (i.e., active consideration and systematic deep thinking) is the presence of useful cues for reference. A well-designed reference cue has the potential to prompt individuals to meticulously reassess their decisions and compare them with the provided references (Weiss 1982). Consequently, individuals can rectify their initial decisions, address conflicts, and even generate novel ideas through cognitive reasoning, association, and imagination (Hollnagel 1987). Several approaches can be effective in fostering such deep thinking. First, high information quality leads to elevated epistemic motivation (Cacioppo et al. 1996). For instance, structured and concrete information can encourage individuals to engage more deeply in a task and, therefore, process information more actively and positively (Mantel and Kardes 1999). Additionally, when individuals are provided with explicit reference points (Chernev 2003), they maintain high motivation to engage in cognitive reasoning and adopt superior information-processing strategies to navigate complex decision-making. Moreover, the decision to employ System 2 processing can be influenced by individuals' experience and expertise. When faced with novel and unfamiliar situations, individuals are more inclined to activate System 2 processing to tackle challenges and gain new knowledge (Smerek 2014).

Applying the lens of this theoretical literature stream to human–machine collaboration, we propose two designs, each of which corresponds to one of the two conditions mentioned above: (1) offering humans and machines rich information for decision-making, and (2) exposing humans to structured machine explanations for final decisions. Specifically, decision-making with rich information requires strong cognitive abilities for information processing (Icard 2018); this arouses humans' perception of the task complexity (Sun and Taylor 2020). We thus posit that, compared with limited information, offering rich information could enhance and maintain humans' awareness of decision-making tasks and their willingness to participate in the tasks, regardless of their capability for handling large information volumes. Furthermore, presenting machines' decisions as recommendations along with proper machine explanations showing how the prediction outcomes were obtained by machines in a faithful and human-interpretable manner (Krishna et al. 2022) can trigger individuals' active cognitive reasoning. For example, if machine explanations are provided, humans can learn from machines' decision-making rationale, trace back their own decision rules, and double-check whether the new knowledge from machines fits and actually improves decision accuracy (Mohseni et al. 2020). We call this the *rethinking* process. In this paper, rethinking or reconsideration refers to the process of carefully reviewing a decision or conclusion that has previously been made to determine whether the initial decision should be changed. It is usually an inquiry into, or reflection on, the most basic given information, or the asking of fundamental questions such as why and how breakthrough improvements were made after observing new signals or outcomes (Jain and Pagrut 2001). Such a self- and system-monitoring process aligns with the concept of active consideration (i.e., Pattern 5) developed by Jussupow et al. (2021), which was concluded to be the best practice for achievement of satisfactory outcomes from human–machine collaboration.

Broadly speaking, notwithstanding the many and broad investigations into human–machine collaboration, there is a dearth of literature unraveling the decision-making process during human interactions with machine assistants under diverse conditions. This paper aims to bridge that void. Particularly, we focus on the role of information complexity and machine explanations in prompting humans to actively rethink and improve the consequent decision outcomes. Given the complex environments covering interactions among information volumes, machine explanations, human experience, and behavioral biases, this question might not have a fixed and intuitive answer. We also reveal the scenarios that can leverage humans' and machines' respective advantages to realize 1 + 1 > 2.

3. Experimentation

3.1. Experimental Background

We partnered with a large Asian microloan company to conduct a field experiment. The microloan company was founded in 2011 and served over 250,000 borrowers by 2018, with unsecured microloans of approximately US$465. The company uses only the owner's money for lending, and their loans are mostly used for temporary financial needs such as supplementary cash flow for small businesses and irregular shopping needs. The loans have a term of 1–7 months and are repaid in monthly installments starting one month after their issuance. The company sets its annual interest rate from 12 to 16%.[1]

To apply for a loan, a borrower is required to provide their basic personal information such as name, phone number, gender, age, educational level, and income level. Subsequently, borrowers are required to choose the loan amount (US$46.5—US$1,240, US$465 by default) and loan term as well as check the annual interest rate. In this study, we focused only on loans with a term of 1, 2, or 3 months.[2] In addition, borrowers are required to clearly state the purpose of the loan. They can then submit their application. Every new application is randomly assigned

[1] The annual interest rate is set on a daily basis, rather than assigning different rates to borrowers with different assessed credit risks. This daily interest rate is generally determined at a mediate level in the online lending market. The company does not announce the actual rates to the market in advance, and borrowers are therefore less likely to decide strategically when to apply to get a lower rate. Such a design allowed us to tease out the potential endogeneity issues brought by interest rates.

[2] This was to facilitate our experimental observation, because it takes a long time to observe and confirm the repayment or default behavior when the loan term is long. We compared the repayment performance among loans of different terms with the historical data and found that the loan term was not highly related to repayment performance.

to a human evaluator who assesses the borrower's credit risk (i.e., default probability) based on the collected information, and makes the final loan-approval decision accordingly. The focal company's loan-approval rate is approximately 47%, similar to the competitors in the market. The main goal of loan screening is to minimize the number of defaulted cases while maintaining the approval rate specified above.

3.2. Experimental Setup

3.2.1 Implementation of Treatment I: Information Complexity

Inspired by the dual-process theories of reasoning, we introduced two factors that could influence human evaluators' decision-making in collaboration with machines in Section 2.4. As the first step, we utilized the focal empirical setup to incorporate variations in information complexity. Before our experiment, the focal platform granted loans based entirely on human evaluators' decisions. Evaluators only accessed borrowers' basic information, loan history, and current loan attributes (12 variables [features] in total) to make their credit risk evaluation. Thus, this information comprises the first level of information complexity: small information volumes. To construct an alternative information scenario (i.e., with large information volumes), we asked the focal company to collect additional information from the borrowers starting June 1, 2017. The additional information included recent (past six months) online shopping activities on the largest e-commerce platform in the focal country and cellphone usage information collected from the pertinent communication carriers.[3] Previous studies have suggested that shopping and cellphone usage may be correlated with borrowers' socioeconomic status and credit behaviors (e.g., Blumenstock et al. 2015). Therefore, based on the relevant literature and canonical behavioral theories (Lu et al. 2023a), we extracted 32 features for each source in order to comprehensively describe borrowers' online shopping and cellphone usage and mobility trace characteristics. Table A1 in Appendix A1 describes these features.

3.2.2 Machine Preparations and Implementation of Treatment II: Machine Explanations

Since the focal company had not sought any machine assistance before our collaboration, it was necessary for us to design and train prediction models for each of the two information scenarios. Our training samples comprised borrowers who submitted loan applications June 1–30, 2017. For these sampled borrowers, the human evaluators

[3] The groups with large information volumes had access to multi-sourced information, emphasizing information diversity (i.e., new attributes). Labeling one treatment as "large information volumes" is intentionally contrasting it with the small-sized demographic feature set used in the other groups. Therefore, our manipulation is intricately aligned with the concept of information complexity, as elucidated in Section 2.4.

assessed their credit risks and made loan-approval decisions using small-scale information, as usual. At this stage, the human evaluators did not have access to the additional information collected. We then gathered repayment information for the approved borrowers from more than 9,000 training sample loans made between July 1 and November 30, 2017. Since the loan term was no longer than 3 months, a 5-month observation period was sufficient for us to confirm borrowers' repayment and default behaviors. Default is defined as the failure to fully repay the loan at least 60 days after the loan due date. At the end of November, we obtained the borrowers' basic and additional information, as well as their repayment behaviors.

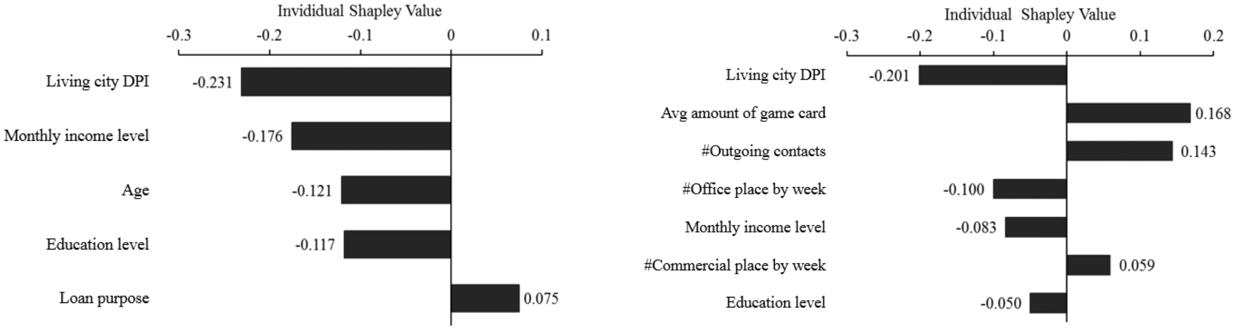
Based on the above information, we then trained machine-learning algorithms. For both information scenarios, we implemented standard operationalizations (e.g., 10-fold cross-validation, out-of-sample prediction, and hyper-parameter tuning) and replicated the training procedures multiple times until they achieved stable loan default prediction performance. We tried diverse, widely accepted machine-learning models, including logistic regression, support vector machine, k-nearest neighbor, multi-level perceptron, random forest, and extreme gradient boosting (XGBoost). XGBoost achieved the best performance, so we employed it in our experiment. To maintain a relatively comparable performance across experimental groups, we did not update XGBoost during the experimental period.

Meanwhile, we leveraged the same training samples to train the human evaluators. Specifically, we randomly separated the human evaluators into two groups: one group maintained the previous loan evaluation process with the small information volume, and the other group evaluated credit risks and made loan-approval decisions with the large information volume. After a 7-day training period, all human evaluators reached a stable evaluation performance. Please refer to Appendix A2 for detailed information on the human evaluators and the training procedure.

With the pre-trained prediction models, we were able to design the second treatment. Specifically, to prepare the machine explanation information based on the above machine-learning algorithms, we implemented a SHAP analysis method, which yields Shapley values representing the average expected marginal contribution to predicting the default probability of one feature after all possible combinations have been considered (Roth 1988). In Figure 1, we present the most important features under the two information volume scenarios.

3.3. Experimental Design

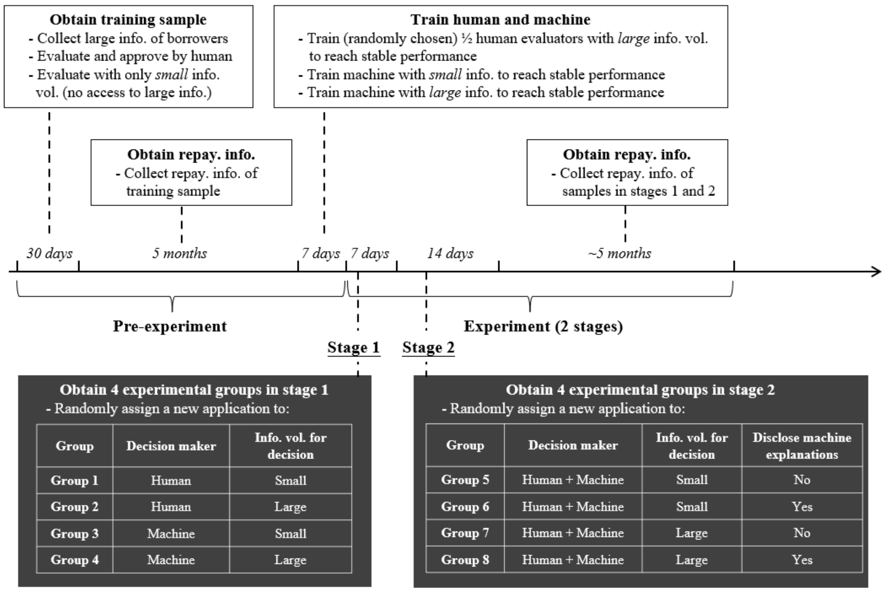
To identify the loan approval decision performance under human-only, machine-only, and human–machine collaboration decision-making scenarios, we designed and implemented a two-stage experiment, as illustrated in Figure 2.



(a) Decisions with Small Information Volume (b) Decisions with Large Information Volume

These features rank in the top 5 or 7 in respective analyses. The other features play only limited roles in machine-learning-based predictions (i.e., they have very small absolute scores). Positive (negative) values mean that the features are positively (negatively) related to default behavior.

Figure 1 Important Features in Machines' Decision-making Processes


Obtain training sample
- Collect large info. of borrowers
- Evaluate and approve by human
- Evaluate with only *small* info. vol. (no access to large info.)
Train human and machine
- Train (randomly chosen) ½ human evaluators with *large* info. vol. to reach stable performance
- Train machine with *small* info. to reach stable performance
- Train machine with *large* info. to reach stable performance
Obtain repay. info.
- Collect repay. info. of training sample
Obtain repay. info.
- Collect repay. info. of samples in stages 1 and 2
30 days
5 months
7 days
7 days
14 days
~5 months
Pre-experiment
Experiment (2 stages)
Stage 1
Stage 2

Obtain 4 experimental groups in stage 1
- Randomly assign a new application to:

Group	Decision maker	Info. vol. for decision
Group 1	Human	Small
Group 2	Human	Large
Group 3	Machine	Small
Group 4	Machine	Large

Obtain 4 experimental groups in stage 2
- Randomly assign a new application to:

Group	Decision maker	Info. vol. for decision	Disclose machine explanations
Group 5	Human + Machine	Small	No
Group 6	Human + Machine	Small	Yes
Group 7	Human + Machine	Large	No
Group 8	Human + Machine	Large	Yes

Figure 2 Experimental Process

Experimental Stage 1. The first stage began on December 8, 2017, and lasted for one week. The relatively short term of the treatments helped tease out the potential confounders stemming from the substantial evolution (learning or change) of the human evaluators, machine-learning algorithms, and borrower-characteristic distributions with long-term experience. At this stage, the company collected basic and additional information from every new borrower, and we randomly assigned the borrowers to one of the four groups. In **Groups 1 (H & S)** and **2 (H & L)**, a credit risk assessment was completed by human evaluators. They had access to the small (Group 1) or large (Group 2) information volumes to inform their approval or rejection of each loan application. The two human evaluator groups were consistent with those in the training process described earlier. In **Groups 3 (M & S)** and **4 (M & L)**, we employed the corresponding pre-trained XGBoost to predict each application's default probability based on a small (Group 3) or large (Group 4) number of features and to make loan-approval decisions by ranking the predicted default probability from lowest to highest. Following the company's usual practice, we maintained the loan-approval rate at 47% in all four experimental groups. For all granted loans, we continued tracing and collecting their repayment behavior from January 8 to May 14, 2018.

Experimental Stage 2. We spent another two weeks (from December 15 to 28, 2017) conducting the second stage of our experiment. The two-week period ensured that the evaluation workload was similar to that in the first stage. Again, we randomly assigned each new loan application to one of the four groups. In all groups, the human evaluators were instructed to collaborate with the machine. Specifically, human evaluators in Group 1 were randomly assigned to Groups 5 and 6 and those in Group 2 were assigned to Groups 7 and 8, with an equal number of evaluators in each group to manage the same amount of information. As illustrated in Figure A2 in Appendix A3, the loan-approval decision process had two steps. In the first step, human evaluators made credit risk evaluation and loan-approval decisions independently with small (Groups 5 and 6) or large (Groups 7 and 8) information volumes; this is identical to the situation in Stage 1. In the second, decision-making step, the machine-learning algorithm's loan-approval decision for the same loan was presented to the human evaluators. In Groups 5 and 6, the machine-learning algorithm used the trained model with a small number of features (corresponding to Group 3), and in Groups 7 and 8, it used a large number of features (corresponding to Group 4). The human evaluators did not have much knowledge of the applied machine-learning algorithm; they were simply notified that machine-learning algorithms usually have strong decision-making abilities.

Next, we incorporated the second treatment, the existence of machine explanations. Specifically, in **Groups 5 ((H + M) & S & w/o Expl)** and **7 ((H + M) & L & w/o Expl)**, we gave only the machine's loan-approval decisions to the human evaluators, without explanations regarding how the decision had been reached (see Figure A2a). In **Groups 6 ((H + M) & S & w/ Expl)** and **8 ((H + M) & L & w/ Expl)**, the human evaluators could see not only the machine's loan-approval decisions but also the post-hoc explanations (i.e., the most important features presented in Figure 1). For these features, the human evaluators could find and compare the values of the fixed features of the focal borrower and the average values of non-defaulters (see Figure A2b). The human evaluators in Groups 6 and 8 were provided this information at the beginning of experimental stage 2. We conjecture, based on our theoretical framework, that this information (strengthened by the value comparison) served as an ideal reference due to machines' superior capability (Chernev 2003). Then, human evaluators were required to make their final loan-approval decisions. When their initial decisions were incongruent with the machine's, they could either insist on their own decisions or adjust them to follow the machine's recommendations. As mentioned before, the human evaluators were told to maintain a consistent approval rate before and during the experiment, and so the approval rates in all of our experimental groups were maintained at approximately 47%. Similarly, we continued to collect the Stage 2 borrowers' repayment performance data over the subsequent 5 months.

3.4. Experimental Data

We obtained our experimental data after completing repayment information collection. The dataset contained the borrowers' basic and additional information, the human evaluators' and machines' initial approval decisions (Groups 1 to 8), the human evaluators' final approval decisions (Groups 5 to 8), and the repayment performance (default or not) of the approved loans. Additionally, we collected background information on the human evaluators, including their gender, education level, number of months' experience (discretized by six month period), and historical decision accuracy (i.e., the ratio of defaulted loans to all approved loans in the three months before our experiment).

Table 1 Randomization Check

		Loan characteristics			Borrower characteristics				
Group	#Obs.	Loan amount (US$)	Interest rate (%)	Loan purpose	Gender	Age	Living city DPI (US$)	Monthly income level	Education level
1. H & S	2,924	472.8	13.888	0.446	0.235	25.17	6,528.9	4.886	4.252
2. H & L	2,930	473.7	13.911	0.437	0.241	25.18	6,505.0	4.886	4.256
3. M & S	3,001	472.9	13.930	0.431	0.249	25.12	6,524.7	4.902	4.201
4. M & L	3,020	472.8	13.913	0.430	0.237	25.19	6,565.2	4.942	4.9206
5. (H + M) & S & w/o Expl	2,885	474.7	13.920	0.437	0.245	25.07	6,545.5	4.960	4.223
6. (H + M) & S & w/ Expl	2,918	470.7	13.902	0.437	0.241	25.15	6,588.1	4.884	4.218
7. (H + M) & L & w/o Expl	2,978	475.2	13.924	0.428	0.233	25.09	6,563.7	4.874	4.216
8. (H + M) & L & w/ Expl	2,946	475.4	13.904	0.434	0.240	25.11	6,571.6	4.943	4.257

Group	#Unique evaluators	Evaluator gender	Evaluator education level	Evaluator months working	Evaluator historical (decision) accuracy
1. H & S	31	0.774	4.452	2.516	2.000
2. H & L	31	0.774	4.452	2.516	2.065

[a] H = human decision, M = machine decision, H + M = human + machine decision, w/o Expl = without AI explanations, w/ Expl = with AI explanations.

[b] Loan purpose: 1 = consumption, 0 = others (e.g., for emergency). Gender: 1 = female, 0 = male.

[c] Monthly income level: 1 = US$150 or below, 2 = US$150–US$300, 3 = US$300–US$450, ..., 8 = US$1,050–US$1,200, 9 = US$1,200 or above.

[d] Education level: 1 = middle school or below, 2 = vocational school, 3 = high school, 4 = technical school, 5 = undergraduate, 6 = graduate or above.

[e] Evaluator months working: 1 = not longer than 6 months, 2 = 6–12 months, 3 = 13–18 months, 4 = longer than 18 months.

[f] Evaluator historical (decision) accuracy: 1 = low (default rate>15%), 2 = medium (10%<default rate<15%), 3 = high (default rate<10%). Refer to Table A2 for descriptive statistics on evaluator historical accuracy.

[g] Groups 3 and 4 did not involve human evaluators. In experimental stage 2, the human evaluators in Group 1 (or 2) were randomly and equally assigned to Groups 5 and 6 (or Groups 7 and 8).

[h] For every feature, the values show no significant differences across the groups based on the *F*-test.

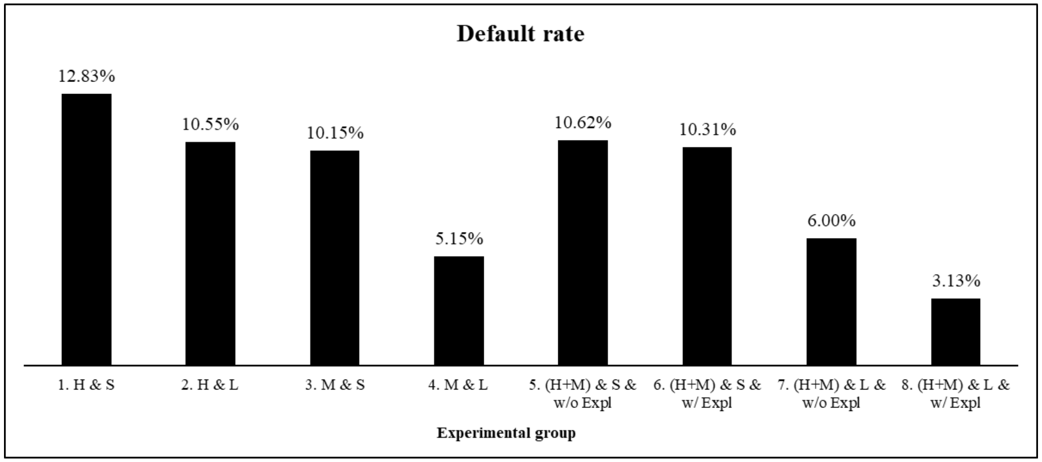


Figure 3 Default Rates of Experimental Groups

There were a total of 23,805 loans in the 8 groups involved in our experiment. We removed 203 repeat borrowers from the company to avoid interference from the previous experience. The final experimental sample size was 23,602. Table 1 reports the sample size and the major characteristics of borrowers, loans, and evaluators across the experimental groups. Most of the borrowers were men (>75%); 28.43% of the borrowers had received an undergraduate education, and the average (self-reported) monthly income ranged between US$450 and US$600. Approximately 44% of the loans were for personal consumption purposes. Regarding the human evaluators, most were female (77%) with a technical school or undergraduate-level education background. On average, the human evaluators had been working for the company for approximately 1 year, and those with high, medium, and low levels of historical decision performances were evenly distributed between the groups (i.e., around $1/3$ each). We detected no statistically significant differences between the groups, which suggested that the randomization had been successful.

4. Empirical Findings

Our key variable of interest was borrowers' default rates. This is a common metric in the microloan industry (Fu et al. 2021) and within the focal company. We defined it as the ratio of defaulted loans to the total number of approved loans. Figure 3 plots the default rates across all groups, and Table 2 calculates the inter-group differences with between-group t tests. The default rate in Group 1 was 12.83%, echoing the average performance of the focal company before our experiment. The comparison yielded several interesting patterns. *First*, as expected, when making decisions separately, the human evaluators performed worse than the machines, and the large-scale information volumes increased the performance gap (i.e., Comparisons B and D). *Second*, the human evaluators did not add additional value when jointly deciding based on a small information volume, regardless of whether the machine explanations were offered (i.e., Comparisons E, G, and H with insignificant differences in the mean value of default rates). *Third*, we observed different outcomes in the scenarios with large information volumes. In particular, when the human evaluators were presented with the machines' suggestions and the machine explanations before making their final decisions, they performed better than the machines' independent decisions, showing a 2.02% reduction in default rates, from 5.15 to 3.13% (i.e., Comparison J). This suggests that the human evaluators contributed additional value to the evaluation process that only they, as humans, could provide. However, this improvement disappeared if no machine explanation was provided (i.e., Comparison I). In sum, the collaborative values were only

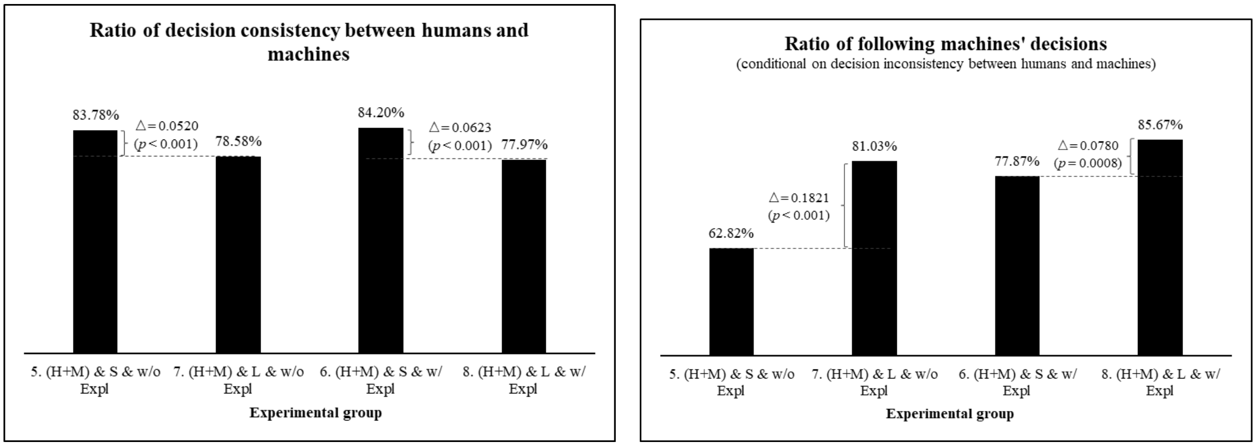
Table 2 Comparison of Default Rates among Different Experimental Groups

Comparison	Experimental groups	Difference in means	p-values
A	1. H & S *vs.* 2. H & L	0.0228	0.0650*
B	1. H & S *vs.* 3. M & S	0.0268	0.0275**
C	3. M & S *vs.* 4. M & L	0.0500	0.0000***
D	2. H & L *vs.* 4. M & L	0.0539	0.0000***
E	5. (H+M) & S & w/o Expl *vs.* 6. (H+M) & S & w/ Expl	0.0032	0.7833
F	7. (H+M) & L & w/o Expl *vs.* 8. (H+M) & L & w/ Expl	0.0287	0.0003***
G	3. M & S *vs.* 5. (H+M) & S & w/o Expl	-0.0048	0.6779
H	3. M & S *vs.* 6. (H+M) & S & w/ Expl	-0.0016	0.8892
I	4. M & L *vs.* 7. (H+M) & L & w/o Expl	-0.0085	0.3215
J	4. M & L *vs.* 8. (H+M) & L & w/ Expl	0.0202	0.0071***

[a] As our experiment comprised multiple treatments, we followed multiple hypothesis testing in experimental economics (List et al. 2019) to address the potential bias. Thus, p-values are multiplicity-adjusted values based on between-group t tests. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

realized if the two conditions, information complexity and useful cues, were satisfied. We also considered profit gains and evaluated the dollar values of the different factors. The results in Figure B1 in Appendix B1 confirmed the consistency.

Noticing the above diverse patterns, we then further decomposed the decision-making behavior of human evaluators after they had observed machines' suggestions. Specifically, in Figure 4, we compared the decision consistency between humans (initial decision) and machines (in Figure 4a), and calculated the adjustment ratios when inconsistency arose (in Figure 4b). Our results indicate that when the human evaluators were making decisions independently (i.e., before observing the machines' suggestions), there were a certain number of cases in which the humans disagreed with the machine's decisions. As shown in Figure 4a, the agreement proportion was smaller with the large information volume (83.78% consistency in Group 5 *vs.* 78.58% consistency in Group 7). This pattern was similar regardless of whether machine explanations were available. In the human–machine collaboration scenario, the human evaluators adjusted their decisions by following the machines' recommendations. The proportion of adjustment, however, varied across the experimental groups. In particular, we observed that only 62.82% disagreement was eliminated with a small information volume and no machine explanation. The adjustment rates significantly increased when the information volume was large or machine explanations were offered. For example, compared with limited information, the availability of large amounts of information could mitigate the human evaluators' unwillingness to follow machines, decreasing it by 18.21% (Group 5 *vs.* Group 7, p-value<0.001). Meanwhile, machine explanations also encouraged human evaluators to accept the machines' decisions by improving the ratio of following from 81.03 to 85.67% (Group 7 *vs.* Group 8, p-value = 0.026).



(a) Ratio of Decision Consistency between Humans and Machines

(b) Ratio of Following Machines' Decisions

Notes: p-values are multiplicity-adjusted values (List et al. 2019) based on between-group t tests.

Figure 4 Consistency and Following between Humans and Machines

5. Mechanism Examinations

This section aims to disentangle the potential mechanisms driving the differences in performance between humans and machines and the contributions made by humans when collaborating with machines. This part consists of three steps. We first examined empirically why humans and machines decided differently when making decisions separately and how decision inconsistency explained the performance differences. Second, we isolated the underlying behavioral mechanisms explaining why humans disagreed with the machines' recommendations when collaboration was allowed. Third, we discussed how disagreement affects decision quality, and decomposed the human evaluators' "rethinking" procedure in the collaborative mode.

5.1. Why Do Humans and Machines Behave Differently?

To answer this question, we explored decision-making processes by identifying the important features involved. First, to determine the information that had played a part in either the human evaluators' or the machines' decision-making processes, we considered a (loan-)application-level Probit model with all available information as independent variables and defined the dependent variable (DV) using a dummy variable, `IfApprove`, which equaled one if the loan was approved. We derived two sets of Probit models using all loan applications with either small or large information volumes. The estimated coefficient of each information variable suggested the predictive power, which served as a proxy for feature importance in the humans' or the machines' decision-making process. Features with significant coefficients in the regressions were important features.

Table 3 **Regressions on Humans' and Machines' Approval Decision (Groups 1, 3, 5, and 6; Probit Model)**

	Groups 3, 5, 6 (machines' decision)		Groups 1, 3, 5, 6 (humans *vs.* machines)	
DV: `IfApprove`	Model 1		Model 2	
Loan purpose	**-0.161*****	**(0.034)**	0.042	(0.048)
Gender	0.030	(0.029)	0.062	(0.058)
Age	**0.087*****	**(0.005)**	0.068	(0.062)
Living city DPI	**0.212*****	**(0.007)**	**0.139*****	**(0.010)**
Monthly income level	**0.126*****	**(0.008)**	**0.082*****	**(0.008)**
Education level	**0.163*****	**(0.021)**	**0.055*****	**(0.028)**
MInd			-0.086	(0.203)
Loan purpose × MInd			**-0.208*****	**(0.040)**
Gender × MInd			-0.030	(0.045)
Age × MInd			**0.024*****	**(0.006)**
Living city DPI × MInd			0.066	(0.060)
Monthly income level × MInd			0.043	(0.036)
Education level × MInd			0.103	(0.087)
Other borrower-related variables	Included		Included	
Other loan-related variables	Included		Included	
Evaluator-related variables	Included		Included	
Log likelihood	-4,951.40		-10,363.18	
#obs.	8,804		17,531	

[a] Model 2 considers human evaluators' initial decisions before displaying machines' recommendations to them when using Groups 5 and 6. We duplicated the sample for Groups 5 and 6 to consider the humans' initial decisions and machines' decisions, respectively.
[b] The variables concretely reported in the table are those that might be useful in this paper's analyses (although they may be insignificant here). Most of the other variables were insignificant, and we do not report their details. Living city DPI was divided by 1,000.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Furthermore, to compare the decision-making processes between humans and machines in a more explicit way, we ran two additional Probit models, in which we included all related loan-level features as well as their interaction terms and a new binary indicator, `MInd`, denoting whether the approval decision was made by a machine learning model ($= 1$ if yes, $= 0$ otherwise). We reported the estimation results in Tables 3 and 4 for the small and large information volumes scenarios. Model 1 in both tables reports the estimates of machine-only decisions. We estimated the coefficients using samples from Groups 3, 5, and 6 for the small information volume scenario and from Groups 4, 7, and 8 for the large information volume scenario. Model 2 in both tables reports the models with interaction terms. We included all human-only decisions (i.e., humans' initial decisions without machine interventions) and machine-only decisions in Groups 1, 3, 5, and 6 (in Table 3) and Groups 2, 4, 7, and 8 (in Table 4). The coefficients of the interaction terms in Model 3 elaborate on whether and to what extent the corresponding features explain the divergence between humans' and machines' decision-making processes.[4]

[4] In Appendix C1, we conducted multiple robustness checks. First, we reran our regressions within each experimental group using different samples. The results indicated that the human evaluators' initial decisions did not involve any learning from the machines' recommendations. This also confirmed that the comparisons between the two stages in our experiment were reasonable. As another robustness check, we employed decision tree approaches to infer the decision rules implemented by human evaluators. The results in Figure C1 confirm the consistency. Additionally, we incorporated an alternative DV to offer more insights into how humans and machines reached the same or different initial decisions.

Table 4 Regressions on Humans' and Machines' Approval Decision (Groups 2, 4, 7, and 8; Probit Model)

	Groups 4, 7, 8 (machines' decision)		Groups 2, 4, 7, 8 (humans *vs.* machines)	
DV: `IfApprove`	Model 1		Model 2	
Loan purpose	**-0.028*****	**(0.004)**	0.021	(0.049)
Gender	0.045	(0.032)	0.070	(0.053)
Age	**0.088*****	**(0.005)**	0.060	(0.074)
Living city DPI	**0.154*****	**(0.007)**	**0.091*****	**(0.011)**
Monthly income level	**0.121*****	**(0.009)**	**0.065*****	**(0.014)**
Education level	**0.075*****	**(0.023)**	**0.072*****	**(0.036)**
Avg amount of game card	**-0.018*****	**(0.001)**	-0.009	(0.016)
ATV shopping durable	0.001	(0.003)	0.005	(0.005)
ATV shopping virtual	-0.001	(0.001)	-0.002	(0.002)
#Outgoing contacts	**-0.052*****	**(0.010)**	**-0.050*****	**(0.018)**
#Office by week	**0.077*****	**(0.003)**	0.004	(0.005)
#Recreational place by week	-0.026	(0.028)	-0.026	(0.042)
#Commercial place by week	**-0.097*****	**(0.008)**	-0.034	(0.069)
#Public service place by week	0.014	(0.014)	0.042	(0.043)
MInd			-0.083	(0.225)
Loan purpose × MInd			**-0.064****	**(0.030)**
Gender × MInd			-0.022	(0.048)
Age × MInd			**0.028*****	**(0.006)**
Living city DPI × MInd			0.056	(0.049)
Monthly income level × MInd			0.006	(0.011)
Education level × MInd			0.006	(0.008)
Avg amount of game card × MInd			**-0.008*****	(0.002)
ATV shopping durable × MInd			0.001	(0.001)
ATV shopping virtual × MInd			-0.001	(0.001)
#Outgoing contacts × MInd			**-0.002***	**(0.001)**
#Office by week × MInd			**0.063*****	**(0.004)**
#Recreational place by week × MInd			0.001	(0.002)
#Commercial place by week × MInd			**-0.063*****	**(0.011)**
#Public service place by week × MInd			-0.027	(0.029)
Other borrower-related variables	Included		Included	
Other loan-related variables	Included		Included	
Evaluator-related variables	Included		Included	
Log likelihood	-4,155.33		-9,642.50	
#obs.	8,944		17,798	

[a] Model 2 considers the human evaluators' initial decisions before the machines' recommendations were displayed to them when using Groups 7 and 8. We duplicated the sample when using Groups 7 and 8 to consider the humans' initial decisions and the machines' decisions, respectively. Other table notes are the same as [b] and [c] in Table 3.

Several interesting patterns explain the differences in performance between the human evaluators' and machines' individual decisions. When decisions were made with the small information volume, the human and machine evaluators considered similar features (i.e., living city DPI, monthly income level, and education level). The machines additionally captured the applicants' age and the loan purpose, which is known to have a relatively high correlation with default behavior (refer to Table 5). This explains why the machines performed slightly better than the humans with the small information volume. When a large information volume was available, the human and machine evaluators deviated. Interestingly, we found that the human evaluators generally tended to stick with traditionally important features (e.g., living city DPI, monthly income level, education level); the only new feature that human evaluators adopted was the frequency of outgoing contacts. In contrast, the machines explored additional sources

Table 5 Correlations of Major Variables

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
(1) IfDefault	1											
(2) Gender	0.025	1										
(3) Living city DPI	**-0.195**	-0.010	1									
(4) Monthly income level	**-0.164**	-0.041	0.022	1								
(5) Age	**-0.120**	-0.054	0.047	**0.104**	1							
(6) Education level	**-0.103**	-0.036	0.002	0.028	**0.091**	1						
(7) Loan purpose	**0.093**	**0.178**	-0.036	-0.033	-0.065	-0.026	1					
(8) Avg amount of game card	**0.169**	**-0.247**	0.001	-0.004	0.014	-0.017	-0.002	1				
(9) #Outgoing contacts	**0.104**	-0.054	-0.023	0.036	0.012	0.033	0.001	-0.027	1			
(10) #Office by week	**-0.115**	-0.013	-0.034	-0.010	-0.021	0.017	0.009	0.046	0.049	1		
(11) #Commercial place by week	**0.090**	**0.096**	-0.030	-0.016	-0.026	0.032	0.014	**-0.089**	0.019	-0.055	1	
(12) ATV shopping virtual	**0.094**	**0.098**	-0.005	-0.070	-0.078	0.035	0.045	-0.038	0.010	-0.028	0.013	1

[a] Correlations are based on all loan samples. Relatively large values are in bold.

of information, with a particular focus on factors potentially linked to default behavior (refer to Table 5). These factors included shopping behavior (e.g., average amounts spent on game cards), cellphone call behavior (e.g., the frequency of outgoing contacts), and offline trajectory behavior (e.g., frequency of visiting the office or commercial places per week). This is reasonable because humans might resist or be incapable of handling new and complicated information (Chapman and Chapman 1967). Moreover, with their increased processing efficiency, machines have been confirmed to have predictive advantages using novel features from alternative data sources (Lu et al. 2023a, Zhou et al. 2021). This also explains the significant improvement achieved by machines with large information volumes.

5.2. Why Do Humans Disagree with Machines' Recommendations?

We next disentangled the underlying behavioral mechanisms when collaboration was employed. We noticed that after observing the machines' recommendations, the human evaluators sometimes adjusted their final decisions to follow the machines' recommendations, but not always. Table 2 shows that only with machine explanations *and* large information volumes did the human evaluators contribute additional value. This value would disappear if either of the two conditions were removed. In order to understand the human evaluators' behavior, we conducted regression tests using observations in which the human evaluators' initial decisions differed from those of the machines.

We employed Probit models, in which the DV was `IfApprove` and the independent variables included all available loan features. To understand how machines' recommendations influenced humans' decision-making processes, we compared the discrepancies between human evaluators' initial and final decisions. Empirically, we defined a new binary indicator, `IfFinal`, which equaled one if the approved decision was made after a machine recommendation was present. Again, we included this binary indicator and the interaction terms of the features to investigate which features played a significant role in changing the human evaluators' decisions.

We reported the results in Tables 6 and 7. With small information volumes (Groups 5 and 6 in Table 6) and with large information volumes but no machine explanations (Group 7 in Table 7), the factors that explained the human evaluators' final approval decisions remained similar to those in the first stage (shown in Tables 3 and 4). For example, with the small information volume, the interaction terms for three features (i.e., applicant age, living city DPI, and monthly income level) presented significant coefficients in Model 2 in Table 3. This suggests that humans rely on these three features to decide whether to change their initial decisions (i.e., whether to follow the machines' recommendations). Take the feature of monthly income level as an example. The corresponding estimate is positive, implying that humans would switch from rejection to approval even if an applicant's income is not high enough. That is, the weight of the monthly income level in human evaluators' credit risk assessment became larger than before, and human evaluators were more tolerant of cases with relatively lower levels of income. To further illustrate humans' willingness to follow machines' suggestions and to capture human behavior in the second stage, we defined another dependent variable, `IfFollow`. The detailed empirical strategy and corresponding results are presented in Appendix C2. All the empirical results imply that, under these experimental conditions, the reasons (i.e., key features) explaining why the human evaluators disagreed with the machines initially were the same as those explaining why they continued to disagree with machines after receiving the machine recommendations.

In Groups 5 to 7, the features with significant estimates of interaction terms with `IfFinal` (i.e., indicating the reasons explaining the differences between initial and final decision-making processes) included both human-familiar ones (i.e., those they had used in the first stage, such as living city DPI and number of outgoing contacts) and machine-only ones (e.g., number of commercial place visits). There are two possible ways that humans and algorithms might reach diverse decisions. One is that humans might have some uncertainty surrounding "borderline" cases (i.e., those with important features showing values near the evaluators' or machines' thresholds). Humans and machines may make inconsistent decisions on such borderline loans when their feature values are located in such threshold gaps. When handling these relatively complicated applications, humans may lack

Table 6 Regression on Human Evaluators' Initial and Final Approval Decisions (Groups 5 and 6; Probit Model)

	Group 5		Group 5		Group 6		Group 6	
	(humans' final decision)		(initial *vs.* final)		(humans' final decision)		(initial *vs.* final)	
DV: `IfApprove`	Model 1		Model 2		Model 3		Model 4	
Loan purpose	-0.025	(0.022)	-0.013	(0.024)	**-0.113***	(0.065)	-0.102	(0.124)
Gender	0.160	(0.140)	0.131	(0.139)	0.037	(0.143)	0.035	(0.144)
Age	**0.073****	**(0.034)**	0.032	(0.040)	**0.070****	**(0.032**	0.024	(0.032)
Living city DPI	**0.155****	**(0.024)**	**0.123*****	**(0.026)**	**0.103*****	**(0.027)**	**0.098*****	**(0.030)**
Monthly income level	**0.117*****	**(0.034)**	**0.096*****	**(0.033)**	**0.059***	**(0.033)**	**0.052***	**(0.028)**
Education level	**0.024*****	**(0.008)**	**0.020****	**(0.008)**	**0.067*****	**(0.015)**	**0.031****	**(0.014)**
IfFinal			**-0.421*****	**(0.085)**			**-0.598*****	**(0.085)**
Loan purpose × IfFinal			-0.012	(0.010)			**-0.011***	(0.006)
Gender × IfFinal			0.028	(0.198)			0.002	(0.183)
Age × IfFinal			**0.041***	**(0.024)**			**0.045****	**(0.023)**
Living city DPI × IfFinal			**0.023***	**(0.013)**			**0.005****	**(0.002)**
Monthly income level × IfFinal			**0.022****	**(0.010)**			**0.005***	**(0.003)**
Education level × IfFinal			0.004	(0.003)			**0.035****	**(0.017)**
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-306.15		-599.31		-293.04		-584.36	
#obs.	468		936		461		922	

[a] Models 1 to 4 are based on the samples in which human evaluators' initial decisions were inconsistent with machines' decisions (i.e., `IfConsistent` $= 0$). We duplicated the sample because we considered the humans' initial and final decisions separately. Other table notes are the same as [b] and [c] in Table 3.

confidence (Kunimoto et al. 2001) and be more likely to follow machines' suggestions, regardless of their initial approval or rejection decisions. Considering the following ratios and the performance improvement from Group 1 to Groups 5 and 6 and from Group 2 to Group 7, our findings indicate that the machines were relatively better at evaluating cases with feature values near the borderline.

Conversely, it is likely that humans and machines could reach distinct conclusions about an applicant's default probability because of differences in evaluating important features. As a result, humans would tend to stick with their initial opinions. Considering that the machines incorporated extra features to assess the loans, these additional features might dominate the human-familiar ones, and human evaluators could find that the values of their familiar features were beyond their expectations. This echoes the literature about humans' aversion toward AI when humans cannot successfully interpret the reasoning behind a machine's decision (Wang and Benbasat 2016). Figure 5 provides empirical evidence with feature distributions to support these arguments. In specific, we visualized the distributions using four sub-samples, which were separated by two standards: whether human evaluators ultimately accepted or rejected the applications (A vs. R), and whether humans followed or continued to disagree with the machines' recommendations (F vs. D). Interestingly, we observed that the means of (F&A) are close to those of (F&R), implying that when dealing with borderline cases, humans place more trust in the machines. On the

Table 7 Regression on Human Evaluators' Initial and Final Approval Decisions (Groups 7 and 8; Probit Model)

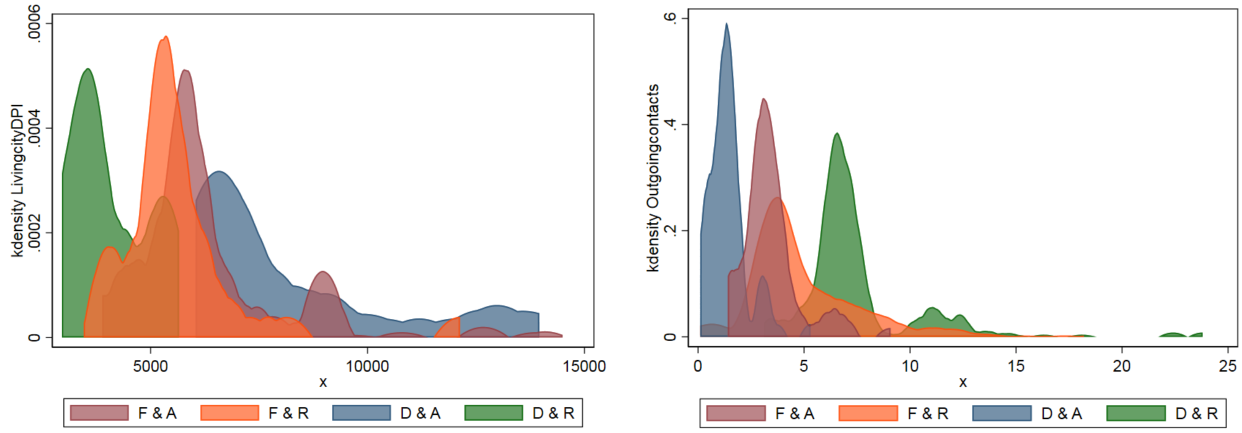
	Group 7 (humans' final decision)		Group 7 (initial *vs.* final)		Group 8 (humans' final decision)		Group 8 (initial *vs.* final)	
DV: `IfApprove`	Model 1		Model 2		Model 3		Model 4	
Loan purpose	-0.059	(0.115)	0.111	(0.117)	-0.017	(0.128)	-0.011	(0.124)
Gender	0.207	(0.136)	0.045	(0.032)	$\mathbf{-0.154}^{**}$	$\mathbf{(0.069)}$	0.032	(0.034)
Age	$\mathbf{0.130}^{**}$	$\mathbf{(0.056)}$	0.073	(0.059)	$\mathbf{0.100}^{***}$	$\mathbf{(0.018)}$	0.051	(0.057)
Living city DPI	$\mathbf{0.135}^{***}$	$\mathbf{(0.025)}$	$\mathbf{0.098}^{***}$	$\mathbf{(0.024)}$	$\mathbf{0.188}^{***}$	$\mathbf{(0.030)}$	$\mathbf{0.140}^{***}$	$\mathbf{(0.033)}$
Monthly income level	$\mathbf{0.076}^{**}$	$\mathbf{(0.032)}$	$\mathbf{0.032}^{***}$	$\mathbf{(0.010)}$	$\mathbf{0.140}^{***}$	$\mathbf{(0.034)}$	$\mathbf{0.110}^{***}$	$\mathbf{(0.033)}$
Education level	$\mathbf{0.191}^{**}$	$\mathbf{(0.081)}$	$\mathbf{0.130}^{***}$	$\mathbf{(0.040)}$	$\mathbf{0.032}^{*}$	$\mathbf{(0.019)}$	$\mathbf{0.030}^{*}$	$\mathbf{(0.016)}$
Avg amount of game card	-0.002	(0.005)	-0.030	(0.057)	-0.038	(0.063)	-0.014	(0.014)
ATV shopping durable	0.003	(0.002)	-0.001	(0.002)	0.007	(0.009)	0.008	(0.012)
ATV shopping virtual	-0.005	(0.004)	-0.001	(0.001)	$\mathbf{-0.020}^{***}$	$\mathbf{(0.005)}$	-0.010	(0.008)
#Outgoing contacts	$\mathbf{-0.036}^{***}$	$\mathbf{(0.010)}$	$\mathbf{-0.015}^{*}$	$\mathbf{(0.008)}$	$\mathbf{-0.025}^{**}$	$\mathbf{(0.012)}$	$\mathbf{-0.022}^{*}$	$\mathbf{(0.012)}$
#Office by week	$\mathbf{0.028}^{**}$	$\mathbf{(0.011)}$	0.067	(0.042)	$\mathbf{0.027}^{**}$	$\mathbf{(0.012)}$	0.019	(0.012)
#Recreational place by week	-0.126	(0.100)	-0.029	(0.095)	-0.034	(0.117)	-0.027	(0.129)
#Commercial place by week	$\mathbf{-0.044}^{*}$	$\mathbf{(0.025)}$	-0.022	(0.024)	$\mathbf{-0.111}^{***}$	$\mathbf{(0.039)}$	-0.058	(0.036)
#Public service place by week	0.064	(0.048)	0.015	(0.048)	0.010	(0.040)	-0.028	(0.040)
IfFinal			$\mathbf{-0.337}^{***}$	$\mathbf{(0.091)}$			$\mathbf{-0.349}^{***}$	$\mathbf{(0.091)}$
Loan purpose × IfFinal			-0.170	(0.164)			-0.006	(0.178)
Gender × IfFinal			0.149	(0.134)			$\mathbf{-0.185}^{*}$	$\mathbf{(0.105)}$
Age × IfFinal			$\mathbf{0.056}^{**}$	$\mathbf{(0.028)}$			$\mathbf{0.048}^{**}$	$\mathbf{(0.025)}$
Living city DPI × IfFinal			$\mathbf{0.053}^{***}$	$\mathbf{(0.014)}$			$\mathbf{0.048}^{***}$	$\mathbf{(0.014)}$
Monthly income level × IfFinal			$\mathbf{0.044}^{***}$	$\mathbf{(0.014)}$			$\mathbf{0.030}^{**}$	$\mathbf{(0.015)}$
Education level × IfFinal			$\mathbf{0.061}^{***}$	$\mathbf{(0.015)}$			0.003	(0.017)
Avg amount of game card × IfFinal			0.027	(0.025)			-0.023	(0.043)
ATV shopping durable × IfFinal			-0.002	(0.004)			-0.001	(0.002)
ATV shopping virtual × IfFinal			-0.004	(0.005)			$\mathbf{-0.010}^{**}$	$\mathbf{(0.004)}$
#Outgoing contacts × IfFinal			$\mathbf{-0.021}^{**}$	$\mathbf{(0.010)}$			-0.003	(0.013)
#Office by week × IfFinal			$\mathbf{0.017}^{*}$	$\mathbf{(0.010)}$			0.008	(0.006)
#Recreational place by week × IfFinal			-0.095	(0.108)			-0.007	(0.114)
#Commercial place by week × IfFinal			$\mathbf{-0.022}^{*}$	$\mathbf{(0.014)}$			$\mathbf{-0.052}^{***}$	$\mathbf{(0.011)}$
#Public service place by week × IfFinal			0.049	(0.050)			0.018	(0.051)
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-329.26		-646.33		-265.37		-546.14	
#obs.	638		1,276		649		1,298	

Table notes are the same as those for Table 6.

contrary, once they found the features were far below or above their thresholds, they held on to their own views. We offer additional evidence to support this assertion by considering all relevant loan features in Figure C2 (Appendix C2).

The above result, however, was not found in Group 8. Interestingly, in Models 3 and 4 of Table 4, we noticed that some alternative features, such as gender and the average transaction amount for purchases of virtual goods (i.e., "ATV shopping virtual"), had significant coefficients. That is, those additional features explained why the human evaluators shifted from their initial decisions.[5] More importantly, given that those features did not reach significance when we compared the human evaluators' initial decisions with those of the machines, it suggests

[5] It is possible that the evaluators might have strategically chosen to follow the machines' decisions if the machine recommended either approval or rejection. We alleviated this concern in Appendix C3.



(a) Distributions of Living City DPI

(b) Distributions of #Outgoing Contacts

[a] All distributions are based on the samples in which human evaluators' initial decisions were inconsistent with machines' decisions.
[b] **F & A**: Cases wherein humans followed the machines' recommendation and ultimately approved the loan applications; **F & R**: Cases wherein humans followed the machines' recommendation and ultimately rejected the loan applications; **D & A**: Cases wherein humans disagreed with the machines' recommendation and ultimately approved the loan applications; **D & R**: Cases wherein humans disagreed with the machines' recommendation and ultimately rejected the loan applications.

Figure 5 Feature Distributions of Diverse Cases (Group 7)

that the human evaluators reconsidered their initial decisions. In other words, the presence of large information volumes and machine explanations provoked evaluators to engage in active rethinking, which improved their final decision accuracy.

5.3. Disagreement and Decision Quality: Decomposition of the Rethinking Process

As discussed earlier, we observed that with the presence of large information volumes and machine explanations, humans reconsidered an interesting feature, "ATV shopping virtual". This feature had not been used by either humans or machines in the independent decision-making process. The prediction models might have ignored or downplayed the values of this feature due to its correlations with other features. We conjectured that the attention to the "ATV shopping virtual" feature stemmed from human evaluators associating it with the "average amount spent on game cards" feature. When the human evaluators saw the machines making different decisions, they also noticed that the loans had some irregular patterns on features that the evaluators were unfamiliar with (e.g., "average amount spent on game cards"). However, such features could hardly be applied by human evaluators, as the most common value by far across all loan applications was 0 (refer to Figure A1a in Appendix A1; the median is 0). Such a distribution would lead to human evaluators perceiving those features as non-informative. The literature has suggested that humans are good at building connections between given information and other relevant, familiar, or understandable information in cognitive processing (Bråten and Samuelstuen 2007, Hollnagel 1987). Since game cards are typical virtual goods and "ATV shopping virtual" had many more salient non-zero values (Figure A1b; the median is 8.70), human evaluators are likely to attend more to this feature when making decisions.

In Appendix C4, we compared the default rates between Groups 7 and 8 after separating loans "saved" by the machines (i.e., those that were originally rejected by human evaluators but ultimately approved due to the machines' approval recommendations) and those "saved" by human evaluators. We showed that using the updated decision rules with new and correct features (i.e., significantly correlated with default behavior), human evaluators were more likely to correctly select "good" loans from those rejected by the machines, whereas humans' decisions to overrule the machines resulted in no change or a decrease in efficiency (i.e., replacing some "bad" applications with other "bad" ones) in Group 7 where humans relied on their priors. Meanwhile, the use of gender features might be due to their relatively high correlations with "ATV shopping virtual" (refer to Table 5). Such findings also explain the alleviation of gender bias (which we will demonstrate later, in Section 6.2). Moreover, we conducted a straightforward post-hoc analysis in Appendix C4 to clarify the allocations of different loan types by humans, machines, and collaborative efforts. This provided additional insights into how machines and humans could assume distinct roles to improve overall collaborative performance.

Taking all of the findings together, our results suggest that with a proper design that invokes humans' active rethinking (e.g., the presence of effective machine explanations when processing complicated information), the collaboration between humans and machines could potentially achieve "1+1>2" in practice. Machines would take responsibility for handling borderline cases, and humans would have the potential to invoke active rethinking to correct machines' mistakes in the "random" cases (e.g., those without explicitly congruent feature patterns) when they perceive that machines have made contradictory decisions, inspired by suggestive information cues.

6. Empirical Extensions

6.1. Heterogeneity by Human Evaluator Characteristics

Recent studies have shown that human agents' degree of decision-making experience might affect their acceptance of machine recommendations as well as their performance in collaboration with machines (Luo et al. 2019, Wang et al. 2023b). Therefore, we decomposed the heterogeneity regarding individual evaluators' characteristics. Below, we focus on the evaluators' experience, based on the length of time (in months) that they had worked in the focal company before we started the experiment. Following Marcotte (1998), the experience was measured at four levels

Table 8 Heterogeneity Analysis of Human Evaluators' Months Working (Probit Model)

	Groups 1 & 2 (only human)		Groups 5–8 (human + machine)					
	Model 1 - *DV:* `IfDefault`		Model 2 - *DV:* `IfDefault`		Model 3 - *DV:* `IfConsistent`		Model 4 - *DV:* `IfFollow`	
Large info. (L)	**-0.384*****	**(0.143)**	**-0.093*****	**(0.018)**	**-0.221*****	(0.083)	**0.469****	**(0.187)**
Month of working=1 (Work=1)	(baseline)		(baseline)		(baseline)		(baseline)	
Work=2	**-0.237***	**(0.127)**	-0.052	(0.157)	0.141	(0.089)	0.141	(0.163)
Work=3	**-0.400*****	**(0.140)**	-0.109	(0.162)	**0.153***	**(0.084)**	-0.212	(0.174)
Work=4	**-0.549***	**(0.146)**	**-0.147***	**(0.087)**	**0.194****	**(0.078)**	-0.284	(0.186)
L × Work=1	(baseline)		(baseline)		(baseline)		(baseline)	
L × Work=2	0.305	(0.187)	-0.103	(0.259)	0.042	(0.112)	0.093	(0.253)
L × Work=3	**0.355***	**(0.212)**	0.383	(0.254)	0.084	(0.120)	0.229	(0.254)
L × Work=4	**0.356***	**(0.203)**	0.057	(0.263)	0.048	(0.113)	0.359	(0.248)
Explanation (Expl)			-0.150	(0.179)	0.114	(0.088)	0.146	(0.192)
Expl × Work=1			(baseline)		(baseline)		(baseline)	
Expl × Work=2			0.405	(0.296)	0.036	(0.113)	0.193	(0.251)
Expl × Work=3			0.057	(0.236)	0.013	(0.120)	**0.720*****	**(0.273)**
Expl × Work=4			-0.021	(0.264)	0.022	(0.122)	**0.442***	**(0.267)**
L × Expl			-0.278	(0.305)	-0.051	(0.121)	0.159	(0.281)
L × Expl × Work=1			(baseline)		(baseline)		(baseline)	
L × Expl × Work=2			-0.196	(0.395)	0.070	(0.158)	-0.480	(0.366)
L × Expl × Work=3			**-0.383****	**(0.180)**	0.059	(0.165)	**-0.641***	**(0.377)**
L × Expl × Work=4			**-0.637***	**(0.375)**	0.078	(0.172)	**-0.998****	**(0.390)**
Borrower-related variables	Included		Included		Included		Included	
Loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-822.87		-5,565.09		-957.51		-1,107.73	
#obs.	2,716		11,727		5,603		2,216	

[a] Models 1 and 2 are based on the approved samples. Model 3 is based on all loan samples. Model 4 is based on the samples in which human evaluators' initial decisions were inconsistent with the machines' decisions (i.e., `IfConsistent` $= 0$). We introduce the definition of `IfFollow` in Appendix C2.

[b] Large info. = 1 for the treatment using large information volumes for decision making, 0 for small. Evaluator months working: 1 = not longer than 6 months, 2 = 6–12 months, 3 = 13–18 months, 4 = longer than 18 months. Interpret. = 1 for treatment of disclosing machine explanations, 0 for not.

[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

(1 = not longer than 6 months, 2 = 6–12 months, 3 = 13–18 months, 4 = longer than 18 months). To quantify the impact of experience levels, we considered another Probit model, this one with three-way interaction terms including the existence of large information volumes, the availability of machine explanations, and experience levels. We also included all lower-level interaction terms in the regression. We presented the estimated coefficients in Table 8, wherein Model 1 considers the default rate as DV and includes humans' independent decisions only, while Models 2–4 are in the human–machine collaboration modes. Specifically, we replicated our mechanism tests with heterogeneous experience levels: whether a loan defaulted was Model 2's DV, whether initial decisions were consistent was Model 3's DV, and whether to follow machines' decisions was Model 4's DV. Note that the estimation of Model 4 incorporated only samples where human evaluators' initial decisions were inconsistent with machine decisions. Additionally, we offer more comprehensive heterogeneity analyses with alternative characteristics in Appendix D1.

Table 8 yields several interesting findings. *First*, the positive estimate of L × Work = 3 (or 4) in Model 1 indicates that without machine assistance, experienced human evaluators performed worse with a large information volume than with a small one. Given the definition of work experience, evaluators with a higher experience level might have accumulated significantly more knowledge in handling small data over a long time, and thus, they might have found it hard to switch their mindset (i.e., experience inertia) (Becker 1995). Another plausible explanation is that these more senior evaluators might have less trust in AI, as suggested by Wang et al. (2023b). On the other hand, evaluators who were new to the company might have still been in the learning stage when the experiment started, and in such cases, persistent learning could have brought more benefits. *Second*, we observed that experienced evaluators tended to make more decisions that were consistent with those of the machines (as shown in the results of Model 3), especially with small information volumes. This is reasonable because experienced evaluators were more likely to have learned the feature values comprehensively and reached a similar level of performance as the machines. *Third*, the estimates in Model 4 suggest that, when we focused on loans with different initial decisions, experienced evaluators were more likely to follow machine explanations in a small information scenario but more likely to overrule machines' decisions and stick to their own opinions given the availability of large information volumes. Combining all of these results with those in Model 2 makes it clear that the satisfaction of both conditions encouraged experienced evaluators to initiate an active rethinking process and thereby achieve reduced borrower credit risk. Furthermore, to deepen our understanding of how individual heterogeneity influences behavior in the presence of machine assistance, we replicated our mechanism examinations with different experience levels. The findings, detailed in Appendix D2, offer more nuanced and straightforward evidence indicating that experienced evaluators were more inclined to initiate an active rethinking process when provided additional external information in Group 8.

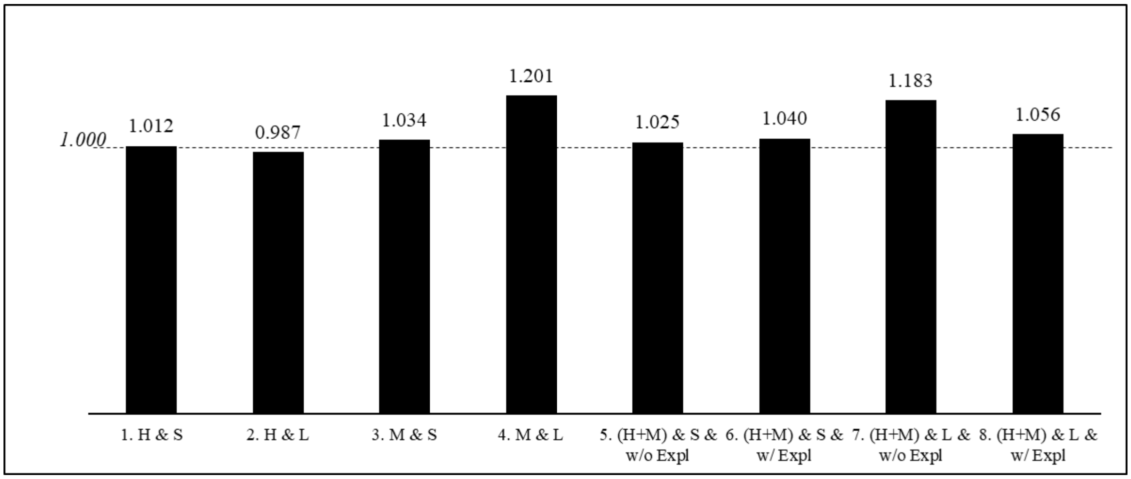
6.2. Decision Biases

As implied in Table 5, most of the major variables considered in both the human evaluators' and machines' decision-making processes were relatively highly correlated with the performance metric. This confirms the fact that both humans and machines made decisions based on their estimated credit risk. In the meantime, it is worth noting that some of the major variables (e.g., loan purpose, average amount spent on game cards, and number of visits to commercial places) were also highly correlated with gender. A natural question arises: will this correlation cause any fairness issues? For example, will it affect the loan-approval decisions of borrowers of different genders, especially when considering different information volumes and human–machine collaboration modes?

To address this question, we first focused on the final performance as measured by the non-default rates. We recorded the statistics of each group in Table D5 in Appendix D3. We observed that with large information volumes (i.e., Group 4), machines tended to favor female applicants, because the non-default rate of the approved male applicants (98.03%) was much larger than that of female applicants (93.99%). That is, machines seemed to have exerted a higher loan-approval criterion for males than females. The involvement of human evaluators without machine explanations (i.e., Group 7) could not alleviate such gender bias. However, when human evaluators were presented with machine explanations (i.e., Group 8), the final repayment performance of the approved female and male applicants became better and similar (96.67% *vs.* 97.55%), suggesting the mitigation of gender bias.

Following Teodorescu et al. (2021), we additionally applied the criterion of "equalized opportunity" (EOR), which requires positive outcomes to be independent of the protected attribute, in order to alternatively measure decision fairness (biases) between genders in our different experimental groups. Let $\mathtt{G}$ be the gender indicator (G = 0 or 1), and $\mathtt{Y=1}$ and $\mathtt{\hat{Y}=1}$ be the correct and actual positive outcomes (i.e., a loan application being approved in our context), respectively. "Correct" here means that non-default loans (observed from the repayment performance of the approved loans) got approved. As such, equalized opportunity means $\mathtt{Pr(\hat{Y}=1|G=0,Y=1)=Pr(\hat{Y}=1|G=1,Y=1)}$. Applying this criterion to our context, EOR describes the decision biases between genders as follows: $\mathtt{EOR} = \frac{\mathtt{Appr(G=0)/NonD(G=0)}}{\mathtt{Appr(G=1)/NonD(G=1)}}$, where $\mathtt{Appr(G=0)}$ and $\mathtt{Appr(G=1)}$ refer to the approval rates for females and males, and $\mathtt{NonD(G=0)}$ and $\mathtt{NonD(G=1)}$ refer to the non-default rates for females and males (calculated within female or male groups), respectively. The closer EOR is to 1, the greater the fairness is between the genders. The larger the deviation from 1, the more bias there is toward females (EOR > 1) or males (EOR < 1). Figure 6 plots the values of EOR across the different experimental groups.

We learned from Figure 6 that the human evaluators treated males and females equally in terms of fairness, regardless of the volumes of information available (EOR = 1.012 (small amounts of information) and 0.987 (large amounts of information), both close to 1). That is, the human evaluators tended to apply relatively similar standards in evaluating the male and female applicants. The machines, however, significantly favored females when they had large information volumes available for decision-making (EOR = 1.201). This was due mainly to the high correlation between the most important features used by machines and the default indicator, as shown in Table 5. This finding is consistent with previous studies (e.g., Fuster et al. 2022) and implies that whereas machines perform much better in general with large-scale information, they return results that are gender-biased, notwithstanding



^{a} Refer to Table D5 for complete values.

Figure 6 Equalized Opportunity Ratio on Gender

the literature's demonstration of the value of large-scale information in alleviating certain forms of demographic discrimination (Lu et al. 2023a). Further, we did not observe any significant change when human evaluators were involved in making the final decisions with small information volumes (i.e., EOR = 1.025 and 1.040 in Groups 5 and 6, respectively). However, we did observe a significant reduction in EOR when both large information volumes and machine explanations were available (i.e., EOR = 1.056 in Group 8). In this scenario, the increase in final decision accuracy could be attributed to human intervention in correcting the risk evaluations of female borrowers. Similarly to our findings in Section 5.2, human evaluators associate certain observed features to others (i.e., "ATV shopping virtual"). Fortunately, the "ATV shopping virtual" feature positively correlates with the feature gender (refer to Table 5) and default probability. Hence, human evaluators helped mitigate gender biases successfully. This again highlights the value and necessity of collaboration between humans and machines. It is essential to acknowledge that the gender bias observed in our dataset and empirical context may be specific to our circumstances. Environments with a more balanced interaction between genders could potentially avoid this gender-related issue. We provided a comprehensive discussion about how our findings concerning gender biases could be extrapolated to other contexts in Appendix D3.

7. Conclusions and Discussion

7.1. Simultaneous Needs of Both Conditions

In the emerging stream of human–machine collaboration literature, there is a dearth of systematic understanding about when, with machines' assistance, humans can actively contribute and how they can add extra value to task outcomes. We dived into the information processing literature, the comprehension of which affords two prerequisites for invoking humans' deep thinking: information complexity initially draws humans' attention to engaging in the tasks, and useful external cues drive humans to perform active consideration. We applied these theoretical implications to human–machine collaboration tasks, and accordingly, against the backdrop of the microloan industry, we devised two treatments by manipulating information volumes and displays of machine explanations. A unique two-stage field experiment helped us to explicitly quantify the corresponding performance.

Our empirical findings shed light on the significance and compatibility of the two theory-driven conditions, and showed that neither can be dispensed with. First, although larger information volumes mean more potential knowledge to help gauge decision-making performance (Hu et al. 2022), our empirical comparisons demonstrated that humans tended to utilize what they have specialized in (i.e., small information volumes, Group 1 *vs.* Group 2), because learning is costly and instant feedback might be uncertain. Without effective extrinsic motivation, such distortion would further impede humans' acceptance of machines' recommendations (Group 4 *vs.* Group 7). This is generally detrimental as humans' insistence on their own decision rules is very likely to result in underfitted decisions in different tasks (Song et al. 2021).

Second, it was also no surprise to find that offering machine explanations alone, without the presence of large information volumes, could not inspire humans' further contribution (Group 6 *vs.* Group 8). This is owed to the fact that machines' superiority in tackling prohibitively (for humans) complex tasks to achieve satisfactory predictions was constrained by information availability (also refer to the comparison between Group 3 *vs.* Group 4 in Figure 3). On top of limited information, humans could not become smarter than machines. Notably, we noticed that a few recent studies have focused on the value of machine explanations to human–machine collaboration (e.g., Bauer et al. 2023, Jacobs et al. 2021). However, our study suggests contingent factors, such as task complexity, would impact the effect of machine explanations. Although machine explanations provide humans with more reference information, humans may not take advantage of them due to insufficient motivation to deeply involve themselves in decision-making (Speier 2006). Instead, humans were found to involve more trust in machines by following their recommendations with those "borderline" loans.

Hence, only with the *simultaneous* presence of large information volumes and machine explanations can human–machine collaboration realize better performance than humans or machines alone (i.e., experimental

Group 8) via initiating active rethinking. This engagement results in further improvement of decision accuracy and mitigation of the machines' biased decisions. Our findings, therefore, confirm the validity of generalizing the dual-process theories of reasoning from humans' independent or interpersonal decision-making to the realm of machine assistance.

7.2. Managerial Implications

This study, built on our unique experimental designs, also offers non-trivial insights to practitioners. Our findings could inform companies' future benefit-cost analyses in managing their efforts/investment and balancing among human agents (human capacities), data purchasing/collecting, and adoption of AI techniques. Our experiments probe diverse possible and manageable factors that could negatively affect the desired efficiency of human–machine collaboration, and we show empirical evidence of those factors' roles in the overall decision-making process. What's more, this paper presents practitioners with a caveat to their prevalent preference for big data, AI techniques, and/or human–machine collaboration. Specifically, if big data is available, this collaboration can achieve both satisfactory decision-making efficiency and fairness. However, when faced with the threat of machines taking their jobs or the possibility of over-domination by machine intelligence, human employees across companies and even industries might resist machine assistance or begin to rely on it excessively. Thus, we provide a scheme of machine interpretability to encourage human agents to rely less on machines and to create additional value. If only small data is available (e.g., affordable), a machine alone seems enough. The involvement of human efforts, regardless of whether machine explanations are present or absent, cannot add significant value in improving prediction accuracy or addressing gender biases in this case. Moreover, our empirical analyses not only offer guidance to platforms in designing efficient collaboration systems but also open pathways to gaining valuable insights into hiring decisions. In particular, our heterogeneity analyses highlight that individuals with experience possess greater potential to attain elevated levels of collaborative performance and amend machine biases through more systematic contributions. Nevertheless, even with experienced employees, platforms should not neglect the importance of refining their training approaches and procedures. This includes the implementation of comprehensive data literacy training programs (Hvalshagen et al. 2023), providing valuable cues and timely feedback for decision-making improvement loops (Proctor and Bonbright 2021). Additionally, it is essential to incorporate modules on ethics and bias awareness into training programs (Sellier et al. 2019).

7.3. Discussions of Generalizability

Our theory-driven experimental design and empirical findings are highly generalizable to other contexts where the decision-making task objectives are not excessively intricate for humans or machines and/or can be clearly formulated. Moreover, the applicability extends to scenarios where opportunities exist to acquire additional information, whether in terms of volume or type, to enhance overall performance (Amit and Sagiv 2013). Examples of such tasks include job candidate screening in labor hiring, supplier evaluation in procurement, and medical treatment decision-making. On a broader note, our study suggests that machines consistently outperform human agents when tasked with objectives that are not particularly challenging, such as classification or prediction problems involving structured objects and features. The availability of a large amount of information might stimulate human agents to pay attention to their tasks, but it does not guarantee that they will aid machines. Additional cues, such as machine explanations, are crucial for guiding human agents to perform an active rethinking of complex information to deal with uncertainties, thereby producing better outcomes.

However, it is worth discussing some caveats to practical system designs as they relate to the generalizability of our results. Our findings regarding the two conditions essential for stimulating active information processing in humans are contingent on many surrounding factors. For example, humans should be responsible for their decision-making performance to some degree, thereby preventing the complete delegation of decision-making to machines. Humans' loan-approval capabilities should also be associated with the ultimate collaboration performance. Also, selected AI algorithms should be suitable for tackling the specific task objectives and models need to be well-trained. Regarding the two focal treatments, rich information is not a panacea; any newly acquired information must be inherently valuable to bolster decision-making performance. In addition, machine explanations must be delivered in a clear and compelling manner. As there might be disagreement between human (expert) knowledge and machine explanations (Krishna et al. 2022), the explanations should be suitably displayed, understandable, and able to stimulate cognitive reasoning (e.g., enabling easy comparisons). Lastly, as proposed by Wang et al. (2023b), human workers' prior knowledge of AI and their responsibilities assigned are factors associated with their attitudes toward AI. In our specific context, human evaluators generally had limited knowledge of machine learning, and the compensation structure within the platform (as outlined in Section 3.1) did not encourage evaluators to proactively enhance their understanding to achieve superior performance levels. In other settings where human workers are more proficient in AI and have stronger motivation to consistently refine their task performance, the responses to machine recommendations (with or without machine explanations) may exhibit variation.

From the technical perspective, our empirical results also reinforce our theory-guided design approach to some extent, as our two proposed treatments did not explicitly rely on any specific form of machine interpretability. As long as they could offer clear signals, we deemed them potentially valuable in encouraging humans to reassess their perspectives. Moreover, while centered on the implementation of specific machine-learning algorithms, our empirical analyses and findings can be extrapolated to diverse applications involving advanced and more intricate AI techniques. On the one hand, our targeted interventions were guided by theory and offered insights into human behavioral responses to factors including task complexity and reference cues. Additionally, our experimental design deliberately withheld information about the specific machine-learning algorithms from participants, making it possible to extend our observations to other AI models, despite potential variations in actual performance and opportunities for human contributions.

Additionally, it is worth noting that in our primary study, we cannot evaluate the value of AI identity explicitly. Put differently, our empirical results do not conclusively discern whether the observed effects stem from the additional information offered by machines (or senior managers) or from the direct attitudes of humans toward AI or machines. However, it is crucial to acknowledge that the performance of senior managers in real-world scenarios may not exhibit the same level of stability and efficiency as machines, especially given the vast amounts of information involved. Considering the time constraints inherent in making accurate decisions, AI or machines tend to outperform their experienced human counterparts. Moreover, several research papers have delved into the difficulties faced by humans when attempting to articulate or summarize the rules guiding their decision-making processes (Hu et al. 2022). Compared to reliance on machines, relying on senior managers to provide explicit decision cues is more challenging. In contrast, machines offer the advantage of leveraging advanced techniques such as feature importance extraction. This underscores the significance of fostering collaboration between humans and machines.

Finally, our experimental design emphasized the efficacy of a two-stage decision-making process wherein where human evaluators initially make independent loan approval decisions and subsequently determine their final decisions by opting to adopt or reject the machine recommendations. While recognizing that two-stage designs may be practically infeasible, we suggest the potential relevance of our findings in scenarios where only a single stage is feasible–directly presenting machine recommendations to the original human-alone decision-making scenario. However, this adjustment may influence decision-making outcomes. For example, without a distinct independent decision-making stage, the direct provision of machine recommendations may lead to over-reliance on machines or foster distrust due to the absence of a clear contrast to humans' independent decisions. The lack of explicit comparisons may further hinder rule identification, especially among less experienced individuals, resulting in more significant heterogeneity in decision-making performance compared to a two-stage setting.

7.4. Limitations and Directions for Future Studies

Our paper has several limitations that provide promising opportunities for future research. *First*, our empirical design focused on a static scenario without human learning. However, in a real-world environment, humans and machines might learn from each other's decision-making processes and adjust gradually over a relatively long period. Future research can extend our analyses to disentangle learning behavior and thereby design an optimization strategy for both sides using techniques such as reinforcement learning models. *Second*, our experimental treatment considered a binary case between small and large information volumes. Future studies can relax this constraint and explore a continuous level of information complexity, the insights from which could offer business managers more practical conclusions and increased value. *Third*, in our empirical setup, we deliberately limited the experimental period to one or two weeks to establish a controlled environment, which helped us mitigate potential biases introduced by human learning behaviors evolving over time. We acknowledge the temporal constraint as a limitation in our study, paving the way for future investigations. Extending the experimental period would enable researchers to explore how humans process and value information conditions over an extended period, offering valuable insights into the dynamics of long-term interactions. *Fourth*, divergence in terms of cultural background or industry domain might have affected our findings. Similar studies in other countries or industries can further validate these findings and offer novel insights into human–machine collaboration designs.

References

Alibaba C (2018) 6 fields where artificial intelligence are surpassing human. *https://www.alibabacloud.com/blog/6-fields-where-artificial-intelligence-are-surpassing-human_584189* .

Allen R, Choudhury P (2022) Algorithm-augmented work and domain experience: The countervailing forces of ability and aversion. *Organization Science* 33(1):149–169.

Amit A, Sagiv L (2013) The role of epistemic motivation in individuals' response to decision complexity. *Organizational Behavior and Human Decision Processes* 121(1):104–117.

Autor DH, Dorn D (2013) How technology wrecks the middle class. *The New York Times* .

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

Bauer K, von Zahn M, Hinz O (2023) Expl (ai) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research* .

Becker HS (1995) The power of inertia. *Qualitative sociology* 18(3):301–309.

Blumenstock J, Cadamuro G, On R (2015) Predicting poverty and wealth from mobile phone metadata. *Science* 350(6264):1073–1076.

Bråten I, Samuelstuen MS (2007) Measuring strategic processing: Comparing task-specific self-reports to traces. *Metacognition and Learning* 2(1):1–20.

Brynjolfsson E, Mitchell T (2017) What can machine learning do? workforce implications. *Science* 358(6370):1530–1534.

Cacioppo JT, Petty RE, Feinstein JA, Jarvis WBG (1996) Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological bulletin* 119(2):197.

Camerer CF, et al. (2019) Artificial intelligence and behavioral economics. *The economics of artificial intelligence: An agenda* 587–608.

Cao S, Jiang W, Wang JL, Yang B (2021) From man vs. machine to man+ machine: The art and ai of stock analyses. Technical report, National Bureau of Economic Research.

Chapman LJ, Chapman JP (1967) Genesis of popular but erroneous psychodiagnostic observations. *journal of Abnormal Psychology* 72(3):193.

Chen D, Li X, Lai F (2017) Gender discrimination in online peer-to-peer credit lending: Evidence from a lending platform in china. *Electronic Commerce Research* 17(4):553–583.

Chen G, Kim KA, Nofsinger JR, Rui OM (2007) Trading performance, disposition effect, overconfidence, representativeness bias, and experience of emerging market investors. *Journal of behavioral decision making* 20(4):425–451.

Chen V, Liao QV, Vaughan JW, Bansal G (2023) Understanding the role of human intuition on reliance in human-ai decision-making with explanations. *arXiv preprint arXiv:2301.07255* .

Chernev A (2003) When more is less and less is more: The role of ideal point availability and assortment in consumer choice. *Journal of Consumer Research* 30(2):170–183.

Choudhury P, Starr E, Agarwal R (2020) Machine learning and human capital complementarities: Experimental evidence on bias mitigation. *Strategic Management Journal* 41(8):1381–1411.

Commerford BP, Dennis SA, Joe JR, Ulla JW (2022) Man versus machine: Complex estimates and auditor reliance on artificial intelligence. *Journal of Accounting Research* 60(1):171–201.

Compeau D, Marcolin B, Kelley H, Higgins C (2012) Research commentary—generalizability of information systems research using student subjects—a reflection on our practices and recommendations for future research. *Information systems research* 23(4):1093–1109.

Davenport T, Guha A, Grewal D, Bressgott T (2020) How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science* 48(1):24–42.

de Véricourt F, Gurkan H (2023) Is your machine better than you? you may never know. *Management Science* .

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Endsley M (1995) Toward a theory of situation awareness in dynamic systems. *Human factors* 37:85–104.

Evans JSB (2003) In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences* 7(10):454–459.

Feuerriegel S, Shrestha YR, von Krogh G, Zhang C (2022) Bringing artificial intelligence to business management. *Nature Machine Intelligence* 4(7):611–613.

Fu R, Huang Y, Singh PV (2021) Crowds, lending, machine, and bias. *Information Systems Research* 32(1):72–92.

Fügener A, Grahl J, Gupta A, Ketter W (2021) Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *Management Information Systems Quarterly (MISQ)-Vol* 45.

Fügener A, Grahl J, Gupta A, Ketter W (2022) Cognitive challenges in human–artificial intelligence collaboration: investigating the path toward productive delegation. *Information Systems Research* 33(2):678–696.

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

Ge R, Zheng Z, Tian X, Liao L (2021) Human–robot interaction: When investors adjust the usage of robo-advisors in peer-to-peer lending. *Information Systems Research* 32(3):774–785.

Germann M, Merkle C (2019) Algorithm aversion in financial investing. *SSRN Electron J* .

Gonzalez L, Loureiro YK (2014) When can a photo increase credit? the impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance* 2:44–58.

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment* 12(1):19.

Guo H, Wang W (2015) An active learning-based svm multi-class classification model. *Pattern recognition* 48(5):1577–1597.

He Y, Xu X, Huang N, Hong Y, Liu D (2020) Preserving user privacy through ephemeral sharing design: A large-scale randomized field experiment in the online dating context. *Available at SSRN 3740782* .

Hollnagel E (1987) Information and reasoning in intelligent decision support systems. *International Journal of Man-Machine Studies* 27(5-6):665–678.

Hu X, Huang Y, Li B, Lu T (2022) Uncovering the source of machine bias. *arXiv preprint arXiv:2201.03092* .

Hu X, Huang Y, Li B, Lu T (2023) Credit risk modeling for financial profitability and fairness: A novel adversarial deep learning model. *Working paper* .

Hvalshagen M, Lukyanenko R, Samuel BM (2023) Empowering users with narratives: Examining the efficacy of narratives for understanding data-oriented conceptual models. *Information Systems Research* 34(3):890–909.

Ibrahim R, Kim SH, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Science* 67(4):2314–2325.

Icard TF (2018) Bayes, bounds, and rational analysis. *Philosophy of Science* 85(1):79–101.

Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ (2021) How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11(1):1–9.

Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624–635.

Jain KJS, Pagrut DS (2001) Re-engineering the enterprise. *Logistics and Supply Chain Management* 268.

Jussupow E, Spohrer K, Heinzl A, Gawlitza J (2021) Augmenting medical diagnosis decisions? an investigation into physicians' decision-making process with artificial intelligence. *Information Systems Research* 32(3):713–735.

Kahneman D (2011) *Thinking, fast and slow* (macmillan).

Keil M, Tan BC, Wei KK, Saarinen T, Tuunainen V, Wassenaar A (2000) A cross-cultural study on escalation of commitment behavior in software projects. *MIS quarterly* 299–325.

Krishna S, Han T, Gu A, Pombra J, Jabbari S, Wu S, Lakkaraju H (2022) The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602* .

Kunimoto C, Miller J, Pashler H (2001) Confidence and accuracy of near-threshold discrimination responses. *Consciousness and cognition* 10(3):294–340.

Levin IP, Huneke ME, Jasper JD (2000) Information processing at successive stages of decision making: Need for cognition and inclusion–exclusion effects. *Organizational behavior and human decision processes* 82(2):171–193.

Lin M, Viswanathan S (2016) Home bias in online investments: An empirical study of an online crowdfunding market. *Management Science* 62(5):1393–1414.

List JA, Shaikh AM, Xu Y (2019) Multiple hypothesis testing in experimental economics. *Experimental Economics* 22(4):773–793.

Liu M, Tang X, Xia S, Zhang S, Zhu Y, Meng Q (2023) Algorithm aversion: Evidence from ridesharing drivers. *Management Science* .

Lou B, Wu L (2021) Ai on drugs: Can artificial intelligence accelerate drug development? evidence from a large-scale examination of bio-pharma firms. *MIS Quarterly* 45(3).

Loutfi E (2019) What does the future hold for ai-enabled coaching. *Chief Learn. Officer* .

Lu J, Lee D, Kim TW, Danks D (2019) Good explanation for algorithmic transparency. *Available at SSRN 3503603* .

Lu SF, Rui H, Seidmann A (2018) Does technology substitute for nurses? staffing decisions in nursing homes. *Management Science* 64(4):1842–1859.

Lu T, Zhang Y, Li B (2023a) Profit vs. equality? the case of financial risk assessment and a new perspective on alternative data. *MIS Quarterly, Forthcoming* .

Lu X, Huang Y, Zhang Y, Shen L (2023b) Role of presentation explicitness in human–artificial intelligence collaboration: A field study in a loan approval service. *Available at SSRN 4547893* .

Luo X, Qin MS, Fang Z, Qu Z (2021) Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing* 85(2):14–32.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38(6):937–947.

Mantel SP, Kardes FR (1999) The role of direction of comparison, attribute-based processing, and attitude-based processing in consumer preference. *Journal of Consumer Research* 25(4):335–352.

Marcotte DE (1998) The wage premium for job seniority during the 1980s and early 1990s. *Industrial Relations: A Journal of Economy and Society* 37(4):419–439.

Mohseni S, Yang F, Pentyala S, Du M, Liu Y, Lupfer N, Hu X, Ji S, Ragan E (2020) Machine learning explanations to prevent overtrust in fake news detection. *arXiv preprint arXiv:2007.12358* .

Oskamp S (1965) Overconfidence in case-study judgments. *Journal of consulting psychology* 29(3):261.

Peukert C, Sen A, Claussen J (2023) The editor and the algorithm: Recommendation technology in online news. *Management science* .

Proctor A, Bonbright D (2021) Constituent voice: Feedback loops, relationships and continual improvement in complex system change. *Generation Impact: International Perspectives on Impact Accounting*, 53–61 (Emerald Publishing Limited).

Rai A (2020) Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science* 48(1):137–141.

Roth AE (1988) Introduction to the shapley value. *The Shapley value* 1–27.

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1(5):206–215.

Sawyer J (1966) Measurement and prediction, clinical and statistical. *Psychological bulletin* 66(3):178.

Schmidt P, Biessmann F, Teubner T (2020) Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29(4):260–278.

Sellier AL, Scopelliti I, Morewedge CK (2019) Debiasing training improves decision making in the field. *Psychological science* 30(9):1371–1379.

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal* 31(2):47–53.

Smerek RE (2014) Why people think deeply: meta-cognitive cues, task characteristics and thinking dispositions. *Handbook of research methods on intuition* 3–14.

Song QC, Tang C, Wee S (2021) Making sense of model generalizability: A tutorial on cross-validation in r and shiny. *Advances in Methods and Practices in Psychological Science* 4(1):2515245920947067.

Speier C (2006) The influence of information presentation formats on complex task decision-making performance. *International Journal of Human-Computer Studies* 64(11):1115–1131.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2022) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science* 68(2):846–865.

Sun T, Taylor SJ (2020) Displaying things in common to encourage friendship formation: A large randomized field experiment. *Quantitative Marketing and Economics* 18:237–271.

Tao Q, Dong Y, Lin Z (2017) Who can get money? evidence from the chinese peer-to-peer lending platform. *Information Systems Frontiers* 19(3):425–441.

Teodorescu MH, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ml augmentation. *MIS Quarterly* 45(3).

Te'eni D, Yahav I, Zagalsky A, Schwartz D, Silverman G, Cohen D, Mann Y, Lewinsky D (2023) Reciprocal human-machine learning: A theory and an instantiation for the case of message classification. *Management Science* .

Tong S, Jia N, Luo X, Fang Z (2021) The janus face of artificial intelligence feedback: Deployment versus disclosure effects on employee performance. *Strategic Management Journal* 42(9):1600–1631.

Van der Schalk J, Beersma B, Van Kleef GA, De Dreu CK (2010) The more (complex), the better? the influence of epistemic motivation on integrative bargaining in complex negotiation. *European Journal of Social Psychology* 40(2):355–365.

Wang C, Zhang W, Zhao X, Wang J (2019) Soft information in online peer-to-peer lending: Evidence from a leading platform in china. *Electronic Commerce Research and Applications* 36:100873.

Wang L, He Y, Huang N, Liu D, Guo X, Chen G (2023a) The role of ai assistants in livestream selling: Evidence from a randomized field experiment. *Available at SSRN* .

Wang W, Benbasat I (2016) Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *Journal of Management Information Systems* 33(3):744–775.

Wang W, Gao G, Agarwal R (2023b) Friend or foe? teaming between artificial intelligence and workers with variation in experience. *Management Science* .

Wang W, Liu X, Zhang X, Hong Y (2023c) Knowledge trap: Human experts distracted by details when teaming with ai. *Available at SSRN* .

Wang W, Yang M, Sun T (2023d) Human-ai co-creation in product ideation: the dual view of quality and diversity. *Available at SSRN 4668241* .

Weiss JA (1982) Coping with complexity: An experimental study of public policy decision-making. *Journal of Policy Analysis and Management* 2(1):66–87.

Zhang M, Sun T, Luo L, Golden J (2023) Consumer and ai co-creation: When and why nudging human participation improves ai creation. *Working Paper* .

Zhou J, Wang C, Ren F, Chen G (2021) Inferring multi-stage risk for online consumer credit services: An integrated scheme using data augmentation and model enhancement. *Decision Support Systems* 149:113611.

Online Web Appendix for

$1 + 1 > 2$? Information, Humans, and Machines

Appendix A. Supplementary Experiment and Data Details

A1. Data and Variables

Table A1: Descriptions of Borrower Features

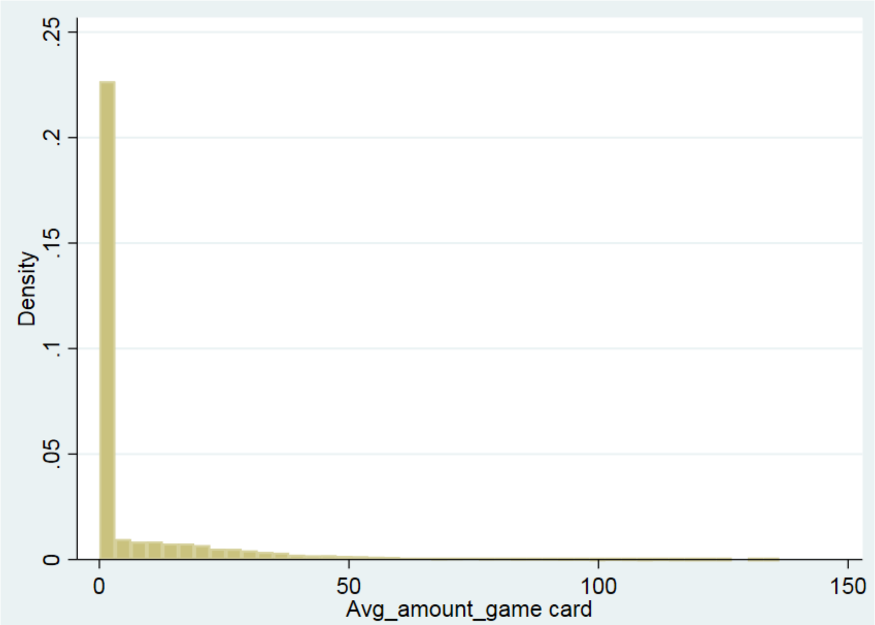
Basic features (Small information volume used for decision making)					
Feature	Description	Mean	S.D.	Min	Max
(1) Gender	Binary; 1 = female, 0 = male	0.24	0.43	0	1
(2) Age	Age	25.13	3.40	18	55
(3) Living city DPI	Disposable personal income (in US$) of a borrower's living city in 2017	6,549.10	2,526.24	3,350.3	13,959.8
(4) Monthly income level	1 = US$150 or below, 2 = US$150–US$300, 3 = US$300–US$450, …, 8 = US$1,050–US$1,200, 9 = US$1,200 or above	4.91	1.89	1	9
(5) Loan-to-income ratio	Ratio of loan amount to monthly income	1.11	1.22	0.03	13
(6) Home ownership	Binary; 1 = self-own, 0 = others	0.17	0.38	0	1
(7) Education level	1 = middle school or below, 2 = vocational school, 3 = high school, 4 = technical school, 5 = undergraduate, 6 = graduate or above	4.23	0.69	1	6
(8) If overdue at other microloan	Binary; 1 = a borrower once became overdue in other microloan platforms, 0 =not	0.08	0.27	0	1
(9) Loan amount	Loan size (in US$)	473.51	78.06	46.5	1240
(10) Loan term	Loan term (in month)	1.91	0.99	1	3
(11) Loan interest rate	Yearly loan interest rate (%)	13.91	0.76	11	24
(12) Loan purpose	Binary; 1 = for (high) consumption (e.g., iPhone and traveling), 0 = for dealing with emergencies (e.g., healthcare, accidents, and business turnover)	0.44	0.50	0	1
Online shopping activity features (Large information volume used for decision making)					
Feature	Description	Mean	S.D.	Min	Max
(13) Amount transfer out	Total amount (in US$) a borrower transferred out on a widely accepted third-party payment platform	548.04	1,683.82	0	57,441.8
(14) Amount transfer in	Total amount (in US$) a borrower transferred in on through a widely accepted third-party payment platform	486.64	1,487.90	0	48,333.3
(15) Ratio transfer out-to-in	=(13)/(14)	1.96	67.94	0	8,783.2
(16) Avg amount of game card	Average amount (in US$) of game card purchase or top up	7.57	16.44	0	136.2
(17) Amount shopping durable	Total amount (in US$) of durable product consumption	481.58	830.93	0	15,876.6
(18) #Order shopping durable	Total number (frequency) of durable product consumption orders	22.75	30.48	0	408
(19) ATV shopping durable	=(17)/(18)	37.65	51.76	0	1,240
(20) #Product shopping durable	Total number of purchased durable products	100.06	1,116.12	0	59,407
(21) Diversity shopping durable	Total number of purchased durable product categories	4.22	2.43	0	9
(22) Amount shopping virtual	Total amount (in US$) of virtual product consumption	389.90	860.24	0	61,436.8
(23) #Order shopping virtual	Total number (frequency) of virtual product consumption orders	30.68	27.81	0	497
(24) ATV shopping virtual	=(22)/(23)	12.79	18.11	0	821.4
(25) #Product shopping virtual	Total number of purchased virtual products	212.44	3,753.50	0	392,612
(26) Ratio shopping amount-to-income	The ratio of average monthly shopping amount to income	0.33	0.91	0	41.0

(27) Variance amount shopping	Standard deviation of weekly amount (in US$) of durable product consumption	119.90	222.61	0	13,602.5
(28) Variance #order shopping	Standard deviation of the weekly number of durable product consumption orders	4.62	4.00	0	96.3
(29) #Order alcohol	Total number of purchased alcohol	0.48	1.39	0	47
(30) #Order caffeine	Total number of purchased caffeine	0.11	0.48	0	13
(31) #Order tobacco	Total number of purchased tobacco	0.00	0.02	0	1
(32) #Order book	Total number of purchased books	0.47	1.38	0	90
(33) #Order medicine	Total number of purchased medicine/drugs	0.21	0.66	0	18
(34) Amount alcohol	Total amount (in US$) of purchased alcohol	17.22	99.92	0	4,060.8
(35) Amount caffeine	Total amount (in US$) of purchased caffeine	2.18	22.59	0	1,177.5
(36) Amount tobacco	Total amount (in US$) of purchased tobacco	0.01	0.47	0	26.7
(37) Amount book	Total amount (in US$) of purchased books	9.00	134.30	0	10,876.1
(38) Amount medicine	Total amount (in US$) of purchased medicine/drugs	2.70	18.47	0	1,034.9
(39) Ratio alcohol to durable amount	=(34)/(17)	0.11	1.80	0	123.2
(40) Ratio caffeine to durable amount	=(35)/(17)	0.00	0.04	0	1.1
(41) Ratio tobacco to durable amount	=(36)/(17)	0.00	0.00	0	0.1
(42) Ratio book to durable amount	=(37)/(17)	0.02	0.17	0	7.7
(43) Ratio medicine to durable amount	=(38)/(17)	0.01	0.10	0	12.0
(44) Ratio shopping for others	Ratio of #orders purchased for others	0.01	0.03	0	1

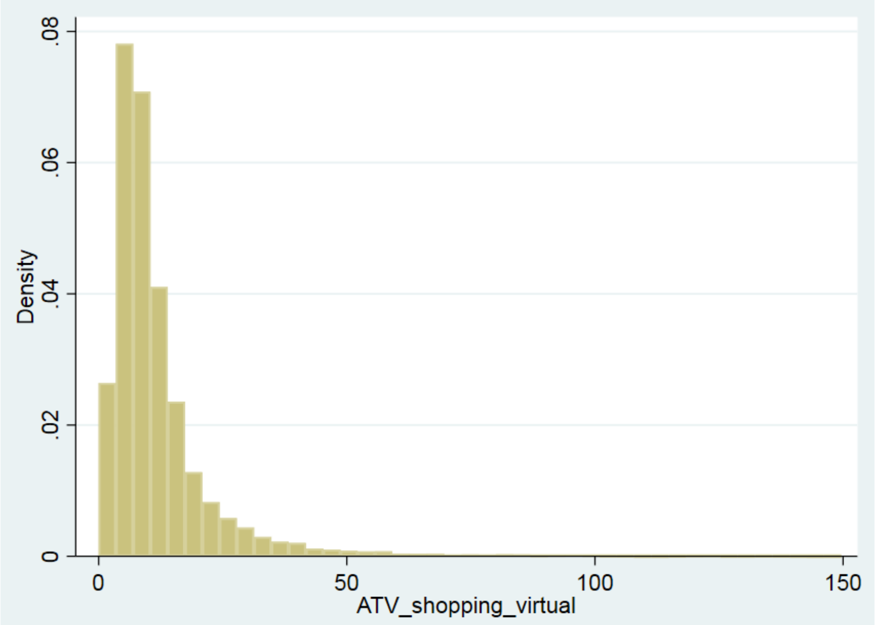
Cellphone usage and mobility trace features (Large information volume used for decision making)

Feature	Description	Mean	S.D.	Min	Max
(45) #Calls by month	=(46)+(48)	178.45	157.12	0	1700
(46) #Calls out by month	Average monthly number of outgoing calls	79.57	80.66	0	1051
(47) #Duration calls out by month	Average monthly duration (in mins) of outgoing calls	125.57	149.72	0	4,019.9
(48) #Calls in by month	Average monthly number of incoming calls	97.21	83.51	0	913.2
(49) #Duration calls in by month	Average monthly duration (in mins) of incoming calls	125.23	120.33	0	2,178.7
(50) Amount phone extra expense	Average monthly extra cellphone expenses (in US$) beyond cellphone plan	1.49	4.99	0	501.0
(51) Ratio #call in-to-out	=(48)/(46)	1.59	1.53	0	41
(52) Ratio duration call in-to-out	=(49)/(47)	1.43	3.09	0	242.1
(53) #Outgoing contacts	Average monthly number of outgoing unique contacted persons	6.38	6.21	0	106.2
(54) #City outgoing contacts	Average monthly number of cities outgoing contacted persons are in	5.12	5.88	0	106.5
(55) #Incoming contacts	Average monthly number of incoming unique contacted persons	6.66	6.50	0	158.1
(56) #City incoming contacts	Average monthly number of cities incoming contacted persons are in	4.87	4.24	0	83.4
(57) #SMS received by month	Average monthly number of short text messages a borrower received	76.03	81.63	1.8	1,533.4
(58) #SMS sent by month	Average monthly number of short text messages a borrower sent	110.39	89.74	9	1,259.9
(59) Ratio SMS received-to-sent	=(57)/(58)	0.73	0.60	0.02	12.8
(60) Cellphone system	Cellphone operation system; 0 = iOS, 1 = Android	0.26	0.44	0	1
(61) Phone number registered duration	Duration (in months) since the cellphone number was registered (started using) by a borrower	42.86	26.38	5	248
(62) #Day phone longest silence	The longest single duration of the cellphone number keeping silent (i.e., no business of calling or messaging happened) in history	5.52	12.77	0	345

(63)	#Data usage by month	Average monthly data usage	807.01	421.63	2.82	2,862.4
(64)	#Defaults first order contacts	Number of loan default a borrower's first-order contacted persons (in recent cellphone call logs) had in the focal and other microloan platforms	0.06	0.39	0	5
(65)	#Apps in cellphone	Number of apps installed in a borrower's cellphone	98.57	39.03	14	199
(66)	#Financial apps	Number of financial and payment apps installed in a borrower's cellphone	4.47	2.07	1	10
(67)	#Using financial apps by week	Average weekly times a borrower used financial and payment apps	3.06	2.10	0	15.1
(68)	#Using social media apps by week	Average weekly times a borrower used social media apps	15.69	9.96	3.1	91.4
(69)	#Using entertainment apps by week	Average weekly times a borrower used entertainment (e.g., video) apps	7.85	4.81	0	30.6
(70)	#Using games apps by week	Average weekly times a borrower used game apps	6.96	6.02	0	26
(71)	#Using news apps by week	Average weekly times a borrower used news apps	8.18	6.66	0	35.3
(72)	#Cities traveled	Total number of cities a borrower appeared in	2.18	1.79	1	16
(73)	#Office by week	Average weekly frequency (times) of appearance in office buildings/areas	16.91	5.56	3.2	34.6
(74)	#Recreational place by week	Average weekly frequency (times) of appearance in entertainment/recreational (e.g., movie theatres and amusement parks) places	0.99	0.81	0	4.9
(75)	#Commercial place by week	Average weekly frequency (times) of appearance in commercial (e.g., shopping malls and restaurants) places	4.46	4.15	0	28.4
(76)	#Public service place by week	Average weekly frequency (times) of appearance in public service (e.g., schools and hospitals) places	3.83	2.67	0	15.1



(a) Distribution of Avg Amount of Game Card (Median: 0)



(b) Distribution of ATV Shopping Virtual (Median: 8.70)

[a] Distributions are based on all loan samples.

Figure A1 Feature Distributions

Table A2 describes human evaluators' characteristics across their historical (decision) accuracy. Generally, male evaluators and evaluators with longer work lengths (e.g., larger than 1 year) have higher decision accuracy.

Table A2 Descriptive Statistics of Human Evaluators on Historical Accuracy

	#Evaluators	Gender	Education level	Months working
Historical accuracy = 1 (low)	20	0.950	4.450	1.900
Historical accuracy = 2 (medium)	21	0.857	4.476	2.714
Historical accuracy = 3 (high)	21	0.524	4.429	2.905
Overall	62	0.774	4.452	2.516

[a] Historical (decision) accuracy: 1 = low (default rate>15%), 2 = medium (10%<default rate<15%), 3 = high (default rate<10%).
[b] Gender: 1 = female, 0 = male.
[c] Education level: 1 = middle school or below, 2 = vocational school, 3 = high school, 4 = technical school, 5 = undergraduate, 6 = graduate or above.
[d] Evaluator months working: 1 = not longer than 6 months, 2 = 6–12 months, 3 = 13–18 months, 4 = longer than 18 months.

A2. Supplementary Experimental Information

In this appendix, we offer supplementary information on the human evaluators in our experiments.

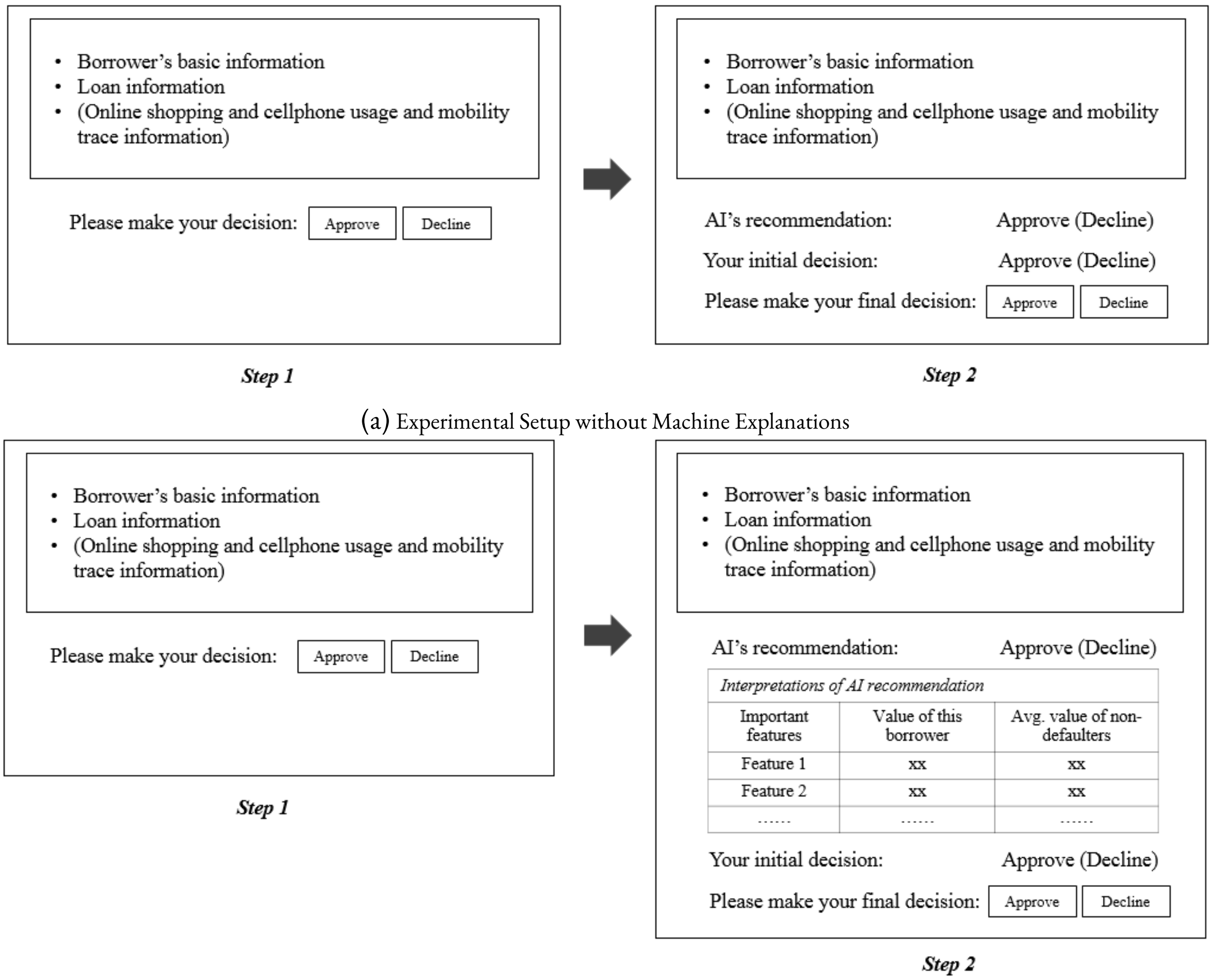
Training Procedure: Before the experiment started, all human evaluators on the focal platforms only had access to the small information volume (i.e., with the predefined 12 features) to evaluate borrowers' credit risks. Leveraging the information collected between June 1 and 30, 2017 (i.e., the same training samples used in the training prediction models), we spent one week training the human evaluators. We randomly separated the human evaluators into two groups: one group maintained the previous loan evaluation process with a small information volume (i.e., basic information), and the other group evaluated credit risks and made loan-approval decisions with a large information volume (i.e., the new sources of information together with the borrowers' basic information; 76 features in total). The human evaluators in the large information volume group were fully familiar with the definition of each feature. We then initiated similar training processes for both sets of human evaluators to help them reach stable evaluator performance. The training process had multiple rounds. In each round, we offered the human evaluators in both groups ten randomly chosen loans and asked them to make loan-approval decisions independently. Then, we disclosed the real repayment performance of these loans to them. After eight rounds, our statistics indicated that the human evaluators reached very stable individual performance in terms of accuracy on default behavior. Importantly, all of the evaluators involved in this experiment had reached stable performance with the small information volume before our experiment, as we did not hire new human evaluators at the beginning of our experiment. The pre-experiment process did not include any additional skill training, except the introduction of large-scale information to one group of evaluators. We ensured that the human evaluators trained with either small or large information volumes remained consistent, with the same level of information complexity throughout the two experimental stages.

Additional Notes on Human Evaluators: In total, we had 62 human evaluators. None of them departed prematurely or were recruited during the course of the experiments. In our experiments, the human evaluators did not know that they were competing with a machine-learning algorithm. The evaluators were informed that AI models typically demonstrate proficiency in decision-making, but they possessed limited knowledge regarding machine learning and the practical applications of AI algorithms when we initiated the experiments.

The roles of the human evaluators in this company are multifaceted. In addition to assessing loan applications, they are assigned responsibilities, including debt collection and various administrative tasks. Delinquent loans are allocated randomly to platform staff for debt collection, and in cases of default, the company may take legal action. Given this context, the human evaluators may lack intrinsic incentives to proactively pursue learning initiatives aimed at achieving enhanced performance levels in the absence of external stimuli.

We acknowledge that machine learning algorithms constitute a subset of AI. However, we emphasize the importance of prediction models in contrast to more intricate AI techniques. These models achieve a balance between accuracy and simplicity, prioritizing interpretability, ease of implementation, and efficiency, particularly in situations where rapid and reliable forecasts are essential. In our specific context, we contend that this implementation is sensible, considering both human comprehension capabilities and the practicality of system implementation.

A3. Experimental Setup



- Borrower's basic information
- Loan information
- (Online shopping and cellphone usage and mobility trace information)

Please make your decision: Approve Decline

Step 1

- Borrower's basic information
- Loan information
- (Online shopping and cellphone usage and mobility trace information)

AI's recommendation: Approve (Decline)

Your initial decision: Approve (Decline)

Please make your final decision: Approve Decline

Step 2

(a) Experimental Setup without Machine Explanations

- Borrower's basic information
- Loan information
- (Online shopping and cellphone usage and mobility trace information)

Please make your decision: Approve Decline

Step 1

- Borrower's basic information
- Loan information
- (Online shopping and cellphone usage and mobility trace information)

AI's recommendation: Approve (Decline)

Interpretations of AI recommendation

Important features	Value of this borrower	Avg. value of non-defaulters
Feature 1	xx	xx
Feature 2	xx	xx
……	……	……

Your initial decision: Approve (Decline)

Please make your final decision: Approve Decline

Step 2

(b) Experimental Setup with Machine Explanations

[a] Experimental Groups 1 and 2 only have the Step 1 page.

[b] Experimental Groups 1, 5, and 6 do not have borrower online shopping and cellphone usage and mobility trace information; Experimental Groups 2, 7, and 8 have such information.

[c] Experimental Groups 5 and 7 do not have machine explanations in Step 2; Experimental Groups 6 and 8 have machine explanations in Step 2.

[d] For Experimental Group 6, the machine explanation table in Step 2 presents the values of the top five most important fixed features extracted from the machine learning training process, based on the full sample. For Experimental Group 8, we show the values of seven fixed features.

The rest of the features are less important, with insignificant or small coefficients. These displayed important features correspond to those in Figure 1.

Figure A2 Experimental Setup

Appendix B. Supplementary Empirical Results

B1. Profit Comparisons

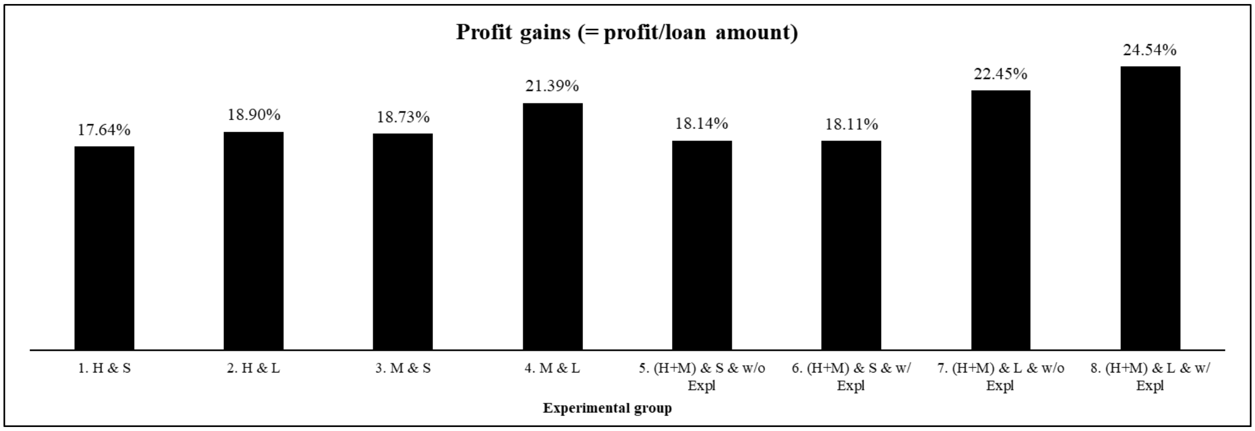


Figure B1 Profits across Experimental Groups

Appendix C. Supplementary Mechanism Examinations

C1. Supplementary Mechanism Analyses: Why Humans and Machines Behave Differently

We fitted Probit models to determine the information that had played a part in human evaluators' decision-making process. The dependent variable was a dummy variable of `IfApprove`, which equaled 1 if human evaluators decided to approve the loans finally and 0 otherwise. We included the 12 traditional features and the 76 variables comprising traditional and alternative features as independent variables, respectively, for the experimental groups wherein human evaluators were offered small information volume (i.e., Groups 1, 5, 6) and large information volume (i.e., Groups 2, 7, 8) for (initial) loan approval decisions. Table C1 and Table C2 report the estimation results, respectively. The statistically significant features were determined to be the important features used in human evaluators' loan approval decisions (Groups 1-2) and initial loan approval decisions made before they had access to the machines' recommendations (Groups 5-8). The consistent patterns among the three models in both tables showed that the human evaluators' initial decisions did not involve any learning from the machines' recommendations. Also, this confirmed that the comparisons between the two stages in our experiment are reasonable. Specifically, Table C1 showed that socioeconomic features such as living city DPI and monthly income level were important in the decision-making processes of all human evaluators when they only had access to a small set of information. When they were exposed to a large set of information (Table C2), they exploited the number of outgoing contacts, another useful feature from the alternative data source, in addition to socioeconomic features.

Table C1 Regressions on Human Evaluators' Approval Decision (Groups 1, 5, and 6; Probit Model)

	Group 1		Group 5		Group 6	
	((only) humans' decision)		(humans' initial decision)		(humans' initial decision)	
DV: `IfApprove`	Model 1		Model 2		Model 3	
Loan purpose	0.027	(0.049)	0.035	(0.049)	0.059	(0.050)
Gender	0.044	(0.057)	0.071	(0.057)	0.063	(0.058)
Age	0.066	(0.065)	0.071	(0.078)	0.069	(0.075)
Living city DPI	**0.126*****	**(0.010)**	**0.130*****	**(0.010)**	**0.160*****	**(0.010)**
Monthly income level	**0.095*****	**(0.013)**	**0.073*****	**(0.013)**	**0.081*****	**(0.013)**
Education level	**0.099*****	**(0.035)**	**0.060***	**(0.035)**	**0.048****	**(0.022)**
Other borrower-related variables	Included		Included		Included	
Other loan-related variables	Included		Included		Included	
Evaluator-related variables	Included		Included		Included	
Log likelihood	-1,809.66		-1,810.12		-1,783.27	
#obs.	2,924		2,885		2,918	

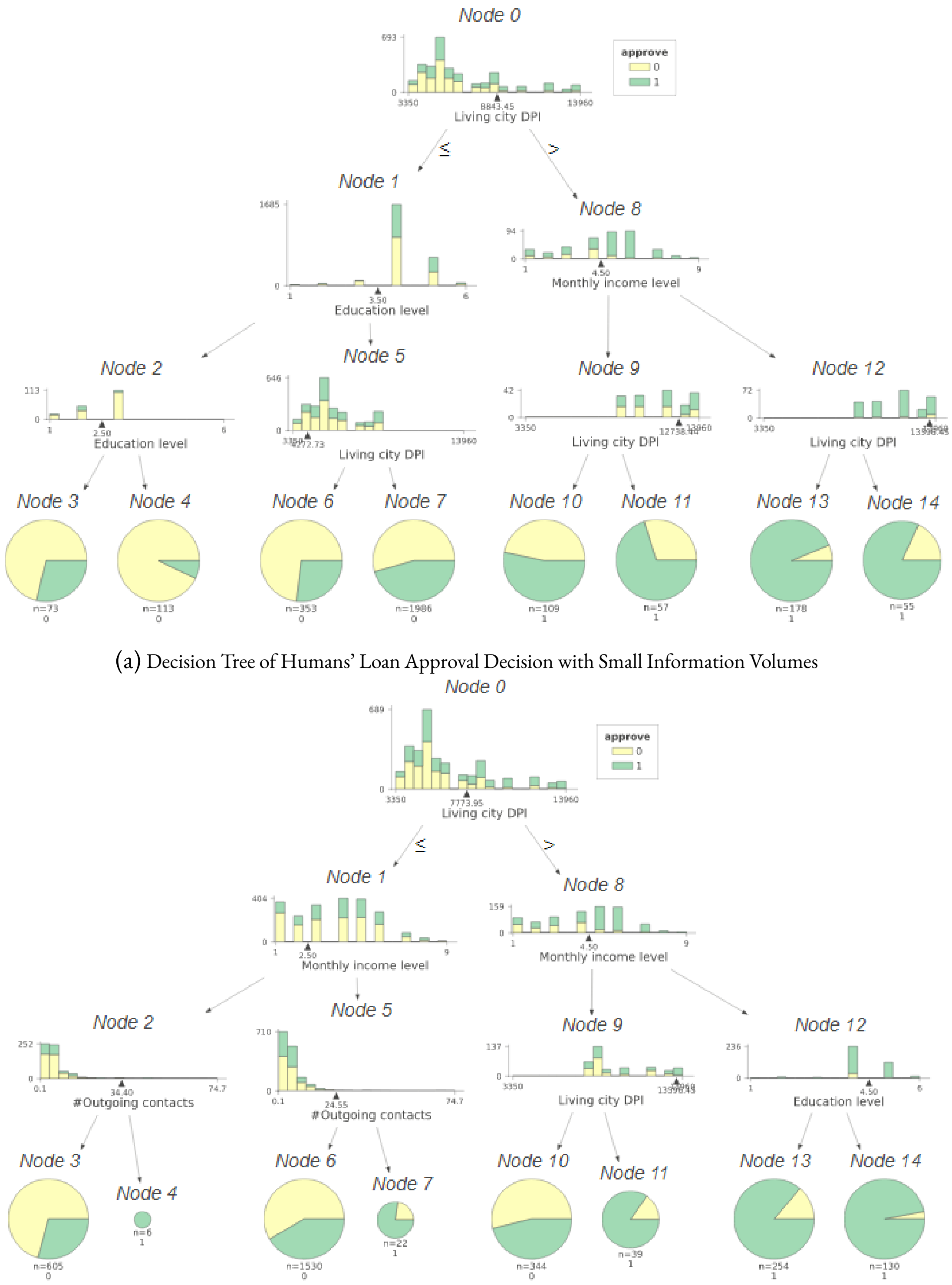
[a] Models 2 and 3 consider human evaluators' initial decisions before the machines' recommendations were displayed to them.
[b] The variables concretely reported in the table are those that might be useful in this paper's analyses (although they may be insignificant here). Most of the other variables are insignificant, and we do not report them. Living city DPI was divided by 1,000.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C2 Regressions on Human Evaluators' Approval Decision (Groups 2, 7, and 8; Probit Model)

	Group 2		Group 7		Group 8	
	((only) humans' decision)		(humans' initial decision)		(humans' initial decision)	
DV: `IfApprove`	Model 1		Model 2		Model 3	
Loan purpose	0.044	(0.050)	0.047	(0.049)	0.040	(0.049)
Gender	0.070	(0.058)	0.093	(0.057)	0.053	(0.057)
Age	0.078	(0.074)	0.050	(0.074)	0.047	(0.074)
Living city DPI	**0.135*****	**(0.011)**	**0.071*****	**(0.010)**	**0.073*****	**(0.010)**
Monthly income level	**0.088*****	**(0.014)**	**0.070*****	**(0.013)**	**0.043*****	**(0.013)**
Education level	**0.073****	**(0.037)**	**0.064***	**(0.036)**	**0.089*****	**(0.036)**
Avg amount of game card	-0.011	(0.017)	-0.007	(0.016)	-0.011	(0.017)
ATV shopping durable	0.004	(0.005)	0.004	(0.006)	0.005	(0.005)
ATV shopping virtual	-0.003	(0.002)	-0.001	(0.002)	-0.003	(0.002)
#Outgoing contacts	**-0.036****	**(0.016)**	**-0.068*****	**(0.015)**	**-0.062*****	**(0.019)**
#Office by week	0.007	(0.005)	0.004	(0.005)	0.002	(0.005)
#Recreational place by week	-0.046	(0.043)	-0.046	(0.044)	-0.021	(0.041)
#Commercial place by week	0.018	(0.013)	0.009	(0.012)	0.085	(0.125)
#Public service place by week	0.048	(0.043)	0.039	(0.031)	0.067	(0.062)
Other borrower-related variables	Included		Included		Included	
Other loan-related variables	Included		Included		Included	
Evaluator-related variables	Included		Included		Included	
Log likelihood	-1,750.93		-1,850.99		-1,798.96	
#obs.	2,930		2,978		2,946	

[a] Models 2 and 3 consider human evaluators' initial decisions before displaying machines' recommendations.
[b] Variables concretely reported in the table are those that might be useful in this paper's analyses (although they may be insignificant here). Most of the other variables are insignificant, and we do not report their details. Living city DPI was divided by 1,000.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Moreover, since understanding humans' decision rules is critical for our empirical conclusions, we further employed a decision tree approach, specifically applying the C4.5 algorithm, using the dependent variable `IfApprove` as the classification label. In decision trees, features closer to the root node usually have greater influences on classification results. The outcomes of the decision tree classification, as depicted in Figure C1, indicated that living city DPI, monthly income level, and education level are important for human evaluators' loan approval decisions in scenarios with small information volumes, while living city DPI, monthly income level, education level, and number of outgoing contacts are crucial in scenarios with large information volumes. Notably, these findings align with the regression analyses presented in Tables 3 and 4 in the main text, showing the robustness of our extracted human decision rules.



(a) Decision Tree of Humans' Loan Approval Decision with Small Information Volumes

(b) Decision Tree of Humans' Loan Approval Decision with Large Information Volumes

Figure C1 Decision Tree of Humans' Loan Approval Decision

To better identify how humans and machines reached different initial decisions, we conducted additional loan-level analyses with a new dependent variable: `IfConsistent`. This variable equals 1 if the human evaluator's initial decision on a given loan is consistent with the machines' recommendation and 0 otherwise. We also employed a Probit model to quantify the effects of different loan-level features driving the consistency. In the regression, we included all of the available loan-level features and evaluator-specific variables. Considering that the coefficient interpretations would be different, we separated the cases based on whether the machine suggested approval or rejection. We defined `MLApprove` as denoting the two cases, with a value of 1 if approved and 0 if rejected).

Features with significant estimates in the regression results are suggested to be important in determining whether humans and machines would reach the same decisions. These significant features could be either those used by both human evaluators and machines or those used by machines only. As this is a relatively complicated interpretation task, we elaborate on the reasoning of this empirical strategy below, using the cases with $\texttt{MLApprove} = 1$ as an example. If features used by both humans and machines are positively correlated with the approval decisions (i.e., negatively correlated with the default probability, referring to Table 5), larger values lead to more consistent decisions, meaning that the coefficient sign in the regression with `IfConsistent` as the DV should be positive as well. Similarly, if a feature negatively correlates with the approval decisions, the corresponding coefficient would also be positive. On the other hand, regarding features used by machines only, we also inferred that when $\texttt{MLApprove} = 1$, the coefficient sign in the regression with `IfConsistent` as the DV would be different from that of the correlation between the focal feature and the approval decisions. To simplify the procedure, we only assumed linear correlations between features and outcomes. This is reasonable because we focused on the perception of human evaluators, who did not impose overly complicated information processing. Furthermore, for features commonly identified by both sides in our context, human evaluators and machines infer the same information (with the same sign). Third, when making decisions separately, human evaluators did not incorporate unique features that were not identified by the machines. Thus, our empirical strategy is feasible.

We report the results in Tables C3 and C4. Overall, the patterns are consistent with our expectations and the previous findings when we regressed `IfApprove` on loan features.

Table C3 Regression Results: Factors and Decision Consistency (Groups 5 and 6; Probit Model)

	Group 5				Group 6			
	Model 1		Model 2		Model 3		Model 4	
DV: `IfConsistent`	`MLApprove` $= 1$		`MLApprove` $= 0$		`MLApprove` $= 1$		`MLApprove` $= 0$	
Loan purpose	0.066	(0.083)	**-0.081***	**(0.047)**	0.043	(0.083)	**-0.088***	**(0.051)**
Gender	-0.057	(0.091)	0.076	(0.097)	-0.020	(0.098)	0.026	(0.092)
Age	**-0.039*****	**(0.012)**	0.021	(0.014)	**-0.027****	**(0.012)**	-0.019	(0.014)
Living city DPI	**0.030****	**(0.015)**	-0.001	(0.022)	**0.074*****	**(0.015)**	-0.015	(0.023)
Monthly income level	**0.048****	**(0.021)**	-0.026	(0.023)	**0.048****	**(0.021)**	-0.016	(0.024)
Education level	**0.117****	**(0.058)**	**-0.134***	**(0.058)**	**0.132****	**(0.060)**	**-0.109***	**(0.056)**
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Other evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-606.53		-655.85		-620.57		-611.35	
#obs.	1,361		1,524		1,389		1,529	

[a] Models 1 to 4 examine the factors related to (in)consistency between human evaluators' initial and machines' loan approval decisions.
[b] Variables concretely reported in the table are those that might be useful in this paper's analyses (though they may be insignificant here). Most of the other variables are insignificant, and thus, we do not report their details. Living city DPI was divided by 1,000.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C4 Regression Results: Factors and Decision Consistency (Groups 7 and 8; Probit Model)

	Group 7				Group 8			
	Model 1		Model 2		Model 3		Model 4	
DV: `IfConsistent`	`MLApprove` $= 1$		`MLApprove` $= 0$		`MLApprove` $= 1$		`MLApprove` $= 0$	
Loan purpose	0.030	(0.183)	**-0.311***	**(0.186)**	0.030	(0.183)	**-0.289***	**(0.174)**
Gender	-0.036	(0.205)	0.047	(0.245)	-0.135	(0.232)	0.246	(0.268)
Age	**-0.083*****	**(0.028)**	0.068	(0.053)	**-0.056****	**(0.026)**	0.050	(0.047)
Living city DPI	**0.086****	**(0.036)**	**-0.130*****	**(0.040)**	**0.110****	**(0.052)**	**-0.164*****	**(0.049)**
Monthly income level	**0.110****	**(0.053)**	**-0.102****	**(0.051)**	**0.114***	**(0.061)**	**-0.113***	**(0.059)**
Education level	**0.086****	**(0.041)**	**-0.172*****	**(0.044)**	**0.088****	**(0.045)**	**-0.109***	**(0.060)**
Avg amount of game card	0.002	(0.005)	**-0.040*****	**(0.012)**	0.004	(0.011)	**-0.041*****	**(0.010)**
ATV shopping durable	0.001	(0.002)	-0.001	(0.003)	0.001	(0.003)	0.010	(0.008)
ATV shopping virtual	-0.002	(0.006)	-0.010	(0.007)	0.009	**(0.008)**	-0.005	(0.011)
#Outgoing contacts	-0.019	(0.017)	0.043	(0.033)	-0.011	(0.018)	0.045	(0.035)
#Office by week	**-0.012***	**(0.007)**	0.026	(0.019)	**-0.033***	**(0.019)**	0.002	(0.022)
#Recreational place by week	0.023	(0.037)	-0.073	(0.050)	0.075	(0.061)	-0.047	(0.062)
#Commercial place by week	**0.441****	**(0.184)**	**-0.293***	**(0.153)**	**0.675****	**(0.262)**	**-0.364****	**(0.181)**
#Public service place by week	0.106	(0.080)	-0.085	(0.077)	0.035	(0.103)	-0.084	(0.082)
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Other evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-762.18		-757.99		-756.26		-734.44	
#obs.	1,402		1,576		1,395		1,551	

Table notes are the same as those for Table C3.

C2. Supplementary Mechanism Analyses: Why Humans Disagree with Machines

In the main analyses, when we decomposed how and why humans disagreed with machines after observing their recommendations, we employed Probit models to identify important features that explained humans' initial and final decisions. In this appendix, we take a further step to directly investigate which features contributed to humans' decisions about following machines and to what degree.

Specifically, we defined a new dependent variable, `IfFollow`, which was equal to 1 if the human evaluator changed his/her mind and followed the machines' decision and was equal to 0 otherwise. We considered a loan-level Probit model by regressing `IfFollow` on all available loan-level features and evaluator-related variables. The coefficients of the loan-level features indicate whether and how much specific features explain humans' final decisions about following the machines' recommendations. Similarly to the analysis in Appendix C1, we separated the cases when $\texttt{MLApprove} = 1$ or 0. We focused on the significant estimates to interpret our findings. We summarized the expected signs of feature coefficients in Table C5.

Table C5 Expected Signs of Feature Coefficients in Tables C6 and C7

Significant Features	`IfApprove`	`IfFollow` ($\texttt{MLApprove} = 1$)	`IfFollow` ($\texttt{MLApprove} = 0$)
Both	+	+	−
	−	−	+
Machine-only	+	−	+
	−	+	−

We reported the estimated results in Tables C6 and C7. In Groups 5, 6, and 7, the significant estimates echo our interpretations in Table C5, implying that humans rely on their standard features to make decisions. That is, the human evaluators' following strategies were influenced by machine-only features when they were not confident with "borderline" cases or human-familiar features when they perceived machines as inscrutable with regard to the feature values. Interestingly, in Group 8, we observed different patterns, with new features emerging as important in shaping humans' thoughts. As we highlighted in the main text, we concluded that humans invoked a systematic "rethinking" process with the help of a high level of information complexity (i.e., large information volumes) and the presence of structured machine explanations.

We conducted principal component analysis (PCA) on several major relevant features–– those identified as significant in Table 7––for loan cases where human evaluators' initial decisions diverged from the machine decisions.

Table C6 Regression on Human Evaluators' Decisions of Following Machines (Groups 5 and 6; Probit Model)

	Group 5				Group 6			
	Model 1		Model 2		Model 3		Model 4	
DV: `IfFollow`	`MLApprove` = 1		`MLApprove` = 0		`MLApprove` = 1		`MLApprove` = 0	
Loan purpose	0.032	(0.186)	**-0.135***	**(0.071)**	0.082	(0.190)	**-0.135***	**(0.076)**
Gender	-0.017	(0.024)	0.026	(0.021)	-0.010	(0.239)	0.011	(0.225)
Age	**-0.063****	**(0.025)**	0.077	(0.059)	**-0.137*****	**(0.038)**	0.052	(0.045)
Living city DPI	0.023	(0.031)	-0.055	(0.042)	0.029	(0.040)	-0.036	(0.047)
Monthly income level	**0.074***	**(0.044)**	-0.081	(0.050)	**0.119****	**(0.055)**	-0.008	(0.055)
Education level	**0.199****	**(0.101)**	**-0.161****	**(0.079)**	**0.127****	**(0.062)**	**-0.052***	**(0.030)**
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Other evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-106.39		-134.02		-97.94		-124.83	
#obs.	229		239		244		217	

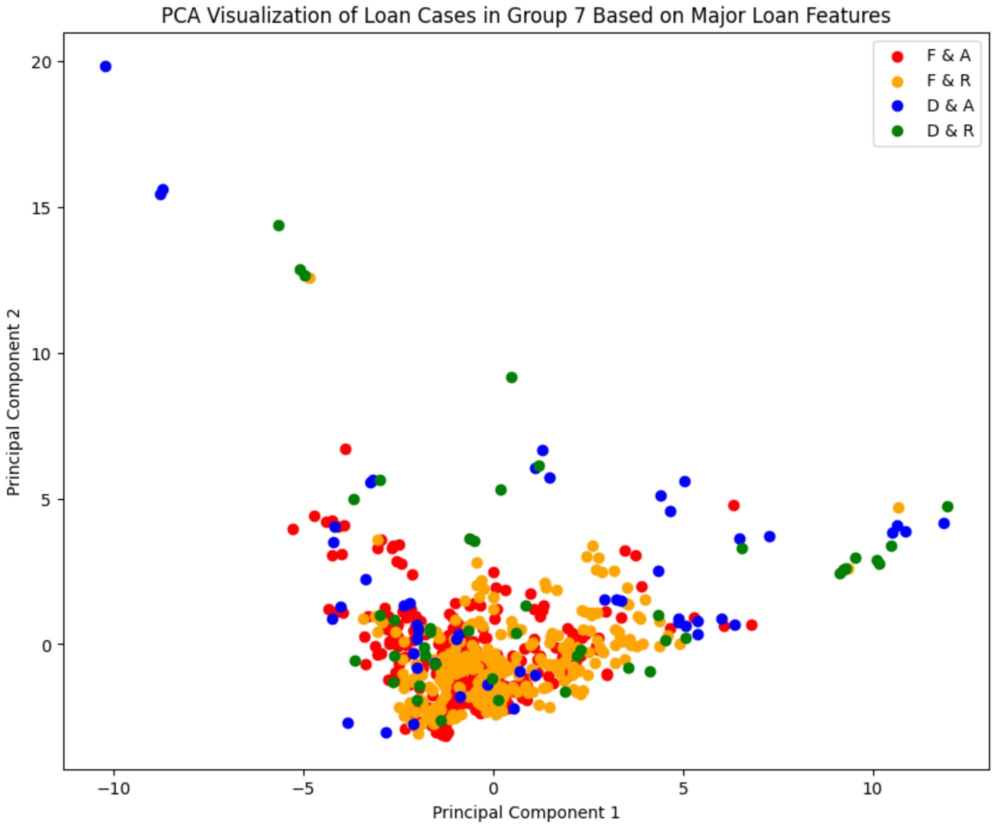
[a] Models 1 to 4 are based on the samples in which human evaluators' initial decisions were inconsistent with machines' decisions (i.e., `IfConsistent` = 0).
[b] Variables concretely reported in the table are those that might be useful in this paper's analyses (although they may be insignificant here). Most other variables are insignificant, and we do not report their details. Living city DPI was divided by 1,000.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table C7 Regression on Human Evaluators' Decisions of Following Machines (Groups 7 and 8; Probit Model)

	Group 7				Group 8			
	Model 1		Model 2		Model 3		Model 4	
DV: `IfFollow`	`MLApprove` = 1		`MLApprove` = 0		`MLApprove` = 1		`MLApprove` = 0	
Loan purpose	0.056	(0.075)	**-0.149****	**(0.075)**	0.119	(0.075)	-0.026	(0.076)
Gender	-0.026	(0.083)	0.053	(0.094)	**-0.094***	**(0.051)**	-0.081	(0.092)
Age	**-0.069*****	**(0.013)**	0.014	(0.011)	**0.060*****	**(0.013)**	0.017	(0.011)
Living city DPI	0.008	(0.013)	-0.028	(0.018)	0.012	(0.013)	0.011	(0.020)
Monthly income level	**0.049****	**(0.020)**	0.004	(0.021)	**0.024****	**(0.020)**	**-0.050****	**(0.022)**
Education level	**0.057****	**(0.025)**	**-0.127****	**(0.052)**	0.014	(0.056)	0.024	(0.055)
Avg amount of game card	0.006	(0.004)	**-0.045****	**(0.030)**	0.002	(0.003)	-0.002	(0.055)
ATV shopping durable	0.001	(0.001)	0.001	(0.001)	-0.001	(0.001)	0.003	(0.003)
ATV shopping virtual	0.001	(0.002)	-0.002	(0.002)	**-0.010****	**(0.004)**	**0.013*****	**(0.003)**
#Outgoing contacts	0.004	(0.005)	0.002	(0.007)	0.006	(0.053)	0.011	(0.008)
#Office by week	**-0.013***	**(0.007)**	0.010	(0.008)	0.001	(0.007)	0.018	(0.027)
#Recreational place by week	0.092	(0.069)	-0.053	(0.065)	0.034	(0.079)	-0.027	(0.059)
#Commercial place by week	**0.030***	**(0.016)**	**-0.043****	**(0.023)**	**-0.072*****	**(0.018)**	**0.032*****	**(0.008)**
#Public service place by week	0.014	(0.032)	-0.001	(0.034)	0.014	(0.034)	0.048	(0.036)
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-123.89		-128.23		-98.39		-102.62	
#obs.	338		300		349		300	

Table notes are the same as those for Table C6.

This analysis aimed to reduce the dimensions of features to two. The resulting plot, presented in Figure C2, revealed insightful patterns. Notably, loans where human evaluators chose to overrule machine recommendations exhibited a more scattered distribution of features (blue and green dots). Conversely, loans where human evaluators chose to follow the machine recommendations had a more concentrated feature distribution (red and orange dots). This observation provides additional support for the assertion that human evaluators are inclined to follow machine recommendations, especially in cases with fewer explicit outlier features, which are often considered "borderline."



[a] Include only the samples in which the human evaluators' initial decisions were inconsistent with the machines' decisions.

[b] **F & A**: Cases wherein humans follow the machines' recommendation and ultimately approve the loan applications; **F & R**: Cases wherein humans follow the machines' recommendation and ultimately reject the loan applications; **D & A**: Cases wherein humans disagree with the machines' recommendation and ultimately approve the loan applications; **D & R**: Cases wherein humans disagree with the machines' recommendation and ultimately reject the loan applications.

Figure C2 PCA Visualization of Loan Cases in Group 7 Based on Major Loan Features

Conversely, when loans exhibit outlier performance in certain key features, human evaluators tend to place greater trust in their own decisions.

C3. Alternative Check

It is possible that evaluators might strategically choose to follow the machines' decisions whenever the machine recommends either approval or rejection. For example, a risk-averse evaluator might tend to follow machines when making rejection decisions. To alleviate concerns that such a preference might have biased our findings, we reran our analyses by including an extra indicator, `MachineApproval`, denoting whether the machine-learning algorithm suggested approval (=1) or rejection (=0). The results shown in Table C8 presented consistent findings, and the extra indicator was not statistically significant.

Table C8 Regression on Human Evaluators' Decisions of Following Machines (Probit Model)

	Group 7		Group 8	
DV: `IfFollow`	Model 1		Model 2	
`MachineApproval`	-0.037	(0.136)	-0.038	(0.150)
Loan purpose	0.011	(0.118)	0.003	(0.126)
Gender	0.039	(0.138)	**-0.089****	**(0.044)**
Age	0.029	(0.020)	**-0.037***	**(0.021)**
Living city DPI	-0.002	(0.023)	0.007	(0.026)
Monthly income level	**0.091*****	**(0.033)**	**0.078****	**(0.034)**
Education level	-0.116	(0.086)	0.131	(0.097)
Avg amount of game card	**-0.022****	**(0.011)**	-0.003	(0.004)
ATV shopping durable	-0.003	(0.002)	0.003	(0.002)
ATV shopping virtual	-0.002	(0.002)	**-0.004***	**(0.002)**
#Outgoing contacts	**0.017****	**(0.009)**	**0.076*****	**(0.013)**
#Office by week	**-0.020***	**(0.012)**	-0.010	(0.012)
#Recreational place by week	0.046	(0.099)	-0.200	(0.114)
#Commercial place by week	**-0.138*****	**(0.025)**	0.013	(0.038)
#Public service place by week	-0.007	(0.049)	0.057	(0.061)
Other borrower-related variables	Included		Included	
Other loan-related variables	Included		Included	
Evaluator-related variables	Included		Included	
Log likelihood	-307.01		-258.32	
#obs.	638		649	

[a] If the machine recommended approval (`MachineApproval`) is a dummy. A value of 1 indicated that the machine recommended approval of the loan, and 0 indicated rejection.
[b] Models 1 and 2 are based on the samples in which the human evaluators' initial decisions were inconsistent with the machines' decisions (i.e., `IfConsistent` $= 0$).
[c] The variables concretely reported in the table are those that might be useful in the analyses (although they may be insignificant here). Most of the other variables were insignificant, and we have not reported their details. Living city DPI was divided by 1,000.
[d] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C4. Disagreement and Decision Quality

Our analyses focused on cases in which the human evaluators' initial decisions were different from the machines' recommendations. Specifically, we first partitioned the loan samples into different cases based on the final decisions of the machines and humans. Table C9 summarizes the numbers of (defaulted) loans in the different cases. In Group 7, when we consider the loan cases wherein the human evaluators overruled the machine decisions (i.e., machines rejected but humans approved), the default rate ($6/59 \approx 10.17\%$) was slightly greater than that for the approved loans with inconsistent initial decisions ($27/335 \approx 8.06\%$). This suggests that without machine explanations, human evaluators may maintain their own opinions wrongly and that, thus, the involvement of human evaluators in the collaborative mode simply replaced some "bad" applications with other "bad" cases. In Group 8, however, we observed a different pattern. We noticed that among all of the 51 loans humans "saved" from the machines' decisions, only one of them defaulted. This supports the value of "rethinking" in human-machine collaborations with machine interpretations available.

Table C9 Decision Making and Default

Case	Humans' *initial* decision	Machine's decision	Humans' *final* decision	Group 7		Group 8	
				#Loans	#Defaulted loans	#Loans	#Defaulted loans
A	Reject	Approve	Approve	276	21	307	1
B	Approve	Reject	Approve	59	6	51	1
A+B	*Approve or Reject*	*Approve or Reject*	*Approve*	*355*	*27*	*358*	*2*
C	Approve	Reject	Reject	241	/	249	/
D	Reject	Approve	Reject	62	/	42	/

[a] The lines in shadow are the cases (B and D) in which human evaluators ultimately disagreed with machines' recommendations when their initial decisions were inconsistent with the machines' decisions (i.e., `IfConsistent` = 0). In our dataset, there was no case wherein the humans' initial decisions and machines' decisions were consistent (either to approve or reject), and the humans changed their minds in the final decisions.
[b] The sign "/" means that the value is unobservable as these loans were rejected by the platform.

Based on the above findings, we conducted post-hoc analyses to elucidate the allocation of various loan types by humans, machines, or collaborative efforts. Identifying the loan categories best suited for each type of evaluator offers valuable guidance for decision-makers aiming to streamline loan processing and optimize overall efficiency. Overall, our empirical findings indicate that machines excel at predicting "borderline" cases, while humans, when provided with appropriate stimuli, can rectify machine errors in extreme cases.

Our post-hoc analyses can be summarized in the following steps, using the observations from Group 7 as an example. Our focal samples consisted of cases where there was a disparity between the humans' initial decisions and the machines' decisions, but the cases eventually received approval and we observed the final outcomes. In other words, we specifically refer to Cases A and B in Table B9. For enhanced interpretability, we also examined scenarios involving only Case A and noted similar patterns.

Step 1: Among all 335 cases with inconsistent decisions between humans and machines, we first needed to identify the "borderline" cases. In specific, we considered four features that were commonly considered by both humans and machines: living city DPI, monthly income level, education level, and number of outgoing contacts (please refer to Table 4 in the paper for references). For each of the four features, we computed the value distributions using all non-defaulted cases.

Step 2: We conducted PCA on these four features to reduce the dimensions of features to two. We delineated "borderline" cases as those with both PCA feature values falling within one standard deviation. We also considered a range of 0.5 standard deviations and obtained relatively consistent results. Subsequently, we observed that the default ratio was 3.77%, indicating that machines demonstrated significantly satisfactory performance in these "borderline" cases.

Step 3: We next evaluated humans' capabilities in processing extreme cases. We defined "extreme" cases as those in which neither of the PCA feature values fell within the specified range in Step 2. This resulted in 154 cases. We categorized these cases into four sub-groups based on the evaluators' degree of work experience and computed the default ratios for each sub-group. The default ratios, from least experienced (i.e., $\texttt{work} = 1$) to most experienced (i.e., $\texttt{work} = 4$) were 13.89%, 15.15%, 12.50%, and 7.55%, respectively. This aligns with our findings that more experienced evaluators tend to contribute to better collaborative performance.

Although the above discussions focused on the large-scale information scenario, our alternative analyses with small-scale information volumes also showed consistency. The key difference in the small-scale information scenario was that one of the four features, the number of outgoing contacts, was replaced with the borrower's age in the PCA. In this sense, we believe our implications can be generalized to situations with limited access to big data.

Appendix D. Supplementary Results of Empirical Extension Analyses

D1. Additional Heterogeneity Analyses

In addition to the work tenure discussed in the paper, we also evaluated the experience-based heterogeneity using an alternative measure: historical decision accuracy (Chen et al. 2007). We reran our analyses using this alternative measure, and the results showed consistency, as indicated in Table D1. Additionally, we considered the heterogeneity of the human evaluators' educational backgrounds. The results in Table D2 do not show significant discrepancies among the evaluators' education levels. A possible reason is the concentrated distribution of education.

Table D1 Heterogeneity Analysis of Human Evaluators' Historical Decision Accuracy (Probit Model)

	Groups 1 & 2 (only human)		Groups 5–8 (human + machine)					
	Model 1 - *DV:* `IfDefault`		Model 2 - *DV:* `IfConsistent`		Model 3 - *DV:* `IfDefault`		Model 4 - *DV:* `IfFollow`	
Large info. (L)	**-0.207***	**(0.115)**	**-0.701*****	**0.064**	**-0.667*****	**(0.139)**	**0.374*****	**(0.132)**
Historical accuracy=1 (Accu=1)	(baseline)		(baseline)		(baseline)		(baseline)	
Accu=2	-0.196	(0.124)	0.049	(0.075)	-0.200	(0.130)	0.037	(0.165)
Accu=3	**-0.733*****	**(0.115)**	**0.114***	**(0.066)**	**-0.272***	**(0.149)**	-0.066	(0.142)
L × Accu=1	(baseline)		(baseline)		(baseline)		(baseline)	
L × Accu=2	-0.154	(0.174)	0.083	(0.097)	-0.020	(0.208)	-0.299	(0.212)
L × Accu=3	**0.515*****	**(0.174)**	-0.063	(0.248)	-0.221	(0.368)	0.268	(0.230)
Explanation (Expl)			0.081	(0.065)	-0.174	(0.125)	0.020	(0.144)
Expl × Accu=1			(baseline)		(baseline)		(baseline)	
Expl × Accu=2			-0.148	(0.105)	0.335	(0.214)	**0.489****	**(0.244)**
Expl × Accu=3			-0.055	(0.092)	0.147	(0.185)	**0.844*****	**(0.210)**
L × Expl			-0.004	(0.094)	-0.160	(0.263)	0.271	(0.201)
L × Expl × Accu=1			(baseline)		(baseline)		(baseline)	
L × Expl × Accu=2			-0.083	(0.139)	-0.460	(0.351)	-0.503	(0.318)
L × Expl × Accu=3			-0.179	(0.137)	**-0.527***	**(0.287)**	**-1.025*****	**(0.319)**
Borrower-related variables	Included		Included		Included		Included	
Loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-806.56		-5,417.54		-955.64		-1,097.68	
#obs.	2,716		11,727		5,603		2,216	

[a] Models 1 and 3 are based on the approved samples. Model 2 is based on all loan samples. Model 4 is based on the samples in which the human evaluators' initial decisions were inconsistent with the machines' decisions (i.e., `IfConsistent` $= 0$).
[b] Large info. = 1 for the treatment using large information volume for decision making, 0 for small. Evaluator historical decision accuracy: 1 = low, 2 = medium, 3 = high. Interpret. = 1 for the treatment of disclosing machine explanations, 0 for not.
[c] Standard errors are in parentheses. Significant results are in bold. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table D2 Heterogeneity Analysis of Human Evaluator Education Levels (Probit Model)

	Groups 1 & 2 (only human)		Groups 5–8 (human + machine)					
	Model 1 - *DV:* `IfDefault`		Model 2 - *DV:* `IfConsistent`		Model 3 - *DV:* `IfDefault`		Model 4 - *DV:* `IfFollow`	
Large info. (L)	**-0.158***	**(0.085)**	**-0.397*****	**(0.151)**	**-0.287***	**(0.168)**	**0.406****	**(0.204)**
Evaluator education=3 (Edu=3)	(baseline)		(baseline)		(baseline)		(baseline)	
Edu=4	0.364	(0.234)	-0.001	(0.125)	-0.030	(0.240)	-0.362	(0.342)
Edu=5	0.202	(0.234)	0.057	(0.122)	-0.154	(0.234)	-0.430	(0.433)
L × Edu=3	(baseline)		(baseline)		(baseline)		(baseline)	
L × Edu=4	-0.279	(0.295)	0.106	(0.163)	-0.516	(0.356)	0.416	(0.408)
L × Edu=5	-0.301	(0.291)	0.279	(0.259)	-0.262	(0.345)	0.410	(0.396)
Explanation (Expl)			0.300	(0.268)	-0.526	(0.342)	0.437	(0.591)
Expl × Edu=3			(baseline)		(baseline)		(baseline)	
Expl × Edu=4			0.274	(0.179)	0.521	(0.364)	-0.115	(0.608)
Expl × Edu=5			0.241	(0.177)	**0.617***	**(0.360)**	0.057	(0.602)
L × Expl			-0.061	(0.208)	-0.124	(0.487)	-0.143	(0.655)
L × Expl × Edu=3			(baseline)		(baseline)		(baseline)	
L × Expl × Edu=4			0.250	(0.225)	-0.050	(0.536)	-0.409	(0.686)
L × Expl × Edu=5			0.265	(0.220)	-0.277	(0.519)	-0.068	(0.673)
Borrower-related variables	Included		Included		Included		Included	
Loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-829.05		-5,613.49		-968.08		-1,106.41	
#obs.	2,716		11,727		5,603		2,216	

[a] Models 1 and 3 are based on the approved samples. Model 2 is based on all loan samples. Model 4 is based on the samples in which the human evaluators' initial decisions were inconsistent with the machines' decisions (i.e., `IfConsistent` $= 0$).

[b] The evaluators' education level has discrete values from 3 to 5. Large info. = 1 for the treatment of using large information volume for decision making, and 0 for small. Interpret. = 1 for the treatment of disclosing machine explanations, 0 for not.

[c] Standard errors are in parentheses. Significant results are in bold. $^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

D2. Supplementary Heterogeneity Analyses for Mechanism Examinations

Table D3 Regression on Human Evaluators' Initial and Final Approval Decisions across Months Working (Group 7; Probit Model)

	Work=1		Work=2		Work=3		Work=4	
	(initial *vs.* final)		(initial *vs.* final)		(initial *vs.* final)		(initial *vs.* final)	
DV: `IfApprove`	Model 1		Model 2		Model 3		Model 4	
Loan purpose	0.077	(0.115)	0.112	(0.118)	0.120	(0.118)	0.106	(0.118)
Gender	0.049	(0.036)	0.048	(0.032)	0.050	(0.036)	0.039	(0.032)
Age	0.094	(0.059)	0.093	(0.059)	0.070	(0.058)	0.065	(0.059)
Living city DPI	**0.099*****	**(0.024)**	**0.111*****	**(0.026)**	**0.078*****	**(0.024)**	**0.100*****	**(0.025)**
Monthly income level	**0.046*****	**(0.013)**	**0.025*****	**(0.010)**	**0.028*****	**(0.010)**	**0.034*****	**(0.010)**
Education level	**0.116*****	**(0.041)**	**0.128*****	**(0.042)**	**0.144*****	**(0.042)**	**0.128*****	**(0.039)**
Avg amount of game card	-0.014	(0.049)	-0.029	(0.050)	-0.032	(0.056)	-0.028	(0.056)
ATV shopping durable	0.001	(0.002)	-0.002	(0.002)	0.000	(0.002)	-0.001	(0.002)
ATV shopping virtual	0.001	(0.001)	-0.001	(0.001)	0.001	(0.001)	-0.001	(0.001)
#Outgoing contacts	-0.013	(0.008)	**-0.019****	**(0.009)**	**-0.016***	**(0.008)**	**-0.016***	**(0.008)**
#Office by week	0.050	(0.041)	0.049	(0.041)	0.054	(0.040)	0.066	(0.040)
#Recreational place by week	-0.04	(0.086)	-0.013	(0.089)	-0.031	(0.100)	-0.028	(0.096)
#Commercial place by week	-0.034	(0.025)	-0.027	(0.025)	-0.010	(0.023)	-0.022	(0.026)
#Public service place by week	0.006	(0.048)	0.015	(0.048)	0.019	(0.048)	0.014	(0.048)
IfFinal	**-0.255*****	**(0.089)**	**-0.346*****	**(0.093)**	**-0.310*****	**(0.090)**	**-0.335*****	**(0.091)**
Loan purpose × IfFinal	-0.082	(0.160)	-0.174	(0.160)	-0.202	(0.158)	-0.144	(0.158)
Gender × IfFinal	0.129	(0.135)	0.149	(0.134)	0.164	(0.135)	0.126	(0.135)
Age × IfFinal	**0.051***	**(0.029)**	**0.052***	**(0.029)**	0.045	(0.029)	**0.060****	**(0.028)**
Living city DPI × IfFinal	**0.047*****	**(0.014)**	**0.031****	**(0.014)**	**0.034****	**(0.014)**	**0.058*****	**(0.014)**
Monthly income level × IfFinal	**0.029****	**(0.014)**	**0.048*****	**(0.016)**	**0.030****	**(0.015)**	**0.040****	**(0.015)**
Education level × IfFinal	**0.068*****	**(0.015)**	**0.051*****	**(0.015)**	**0.068*****	**(0.015)**	**0.045*****	**(0.015)**
Avg amount of game card × IfFinal	0.010	(0.022)	0.033	(0.024)	0.030	(0.026)	0.022	(0.025)
ATV shopping durable × IfFinal	-0.001	(0.004)	-0.004	(0.004)	0.001	(0.004)	-0.001	(0.004)
ATV shopping virtual × IfFinal	-0.001	(0.004)	-0.002	(0.005)	-0.002	(0.005)	-0.004	(0.005)
#Outgoing contacts × IfFinal	-0.010	(0.010)	**-0.018****	**(0.010)**	**-0.021***	**(0.012)**	**-0.024****	**(0.012)**
#Office by week × IfFinal	0.014	(0.010)	0.012	(0.010)	**0.017***	**(0.010)**	**0.018***	**(0.010)**
#Recreational place by week × IfFinal	-0.080	(0.104)	-0.078	(0.108)	-0.101	(0.102)	-0.088	(0.108)
#Commercial place by week × IfFinal	**-0.030****	**(0.014)**	**-0.023***	**(0.014)**	**-0.020***	**(0.010)**	**-0.025****	**(0.012)**
#Public service place by week × IfFinal	0.021	(0.048)	0.046	(0.049)	0.050	(0.046)	0.038	(0.051)
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-204.70		-235.95		-245.31		-299.76	
#obs.	214		242		270		550	

Table notes are the same as those for Table 6.

Table D4 Regression on Human Evaluators' Initial and Final Approval Decisions across Months Working (Group 8; Probit Model)

	Work=1		Work=2		Work=3		Work=4	
	(initial *vs.* final)		(initial *vs.* final)		(initial *vs.* final)		(initial *vs.* final)	
DV: `IfApprove`	Model 1		Model 2		Model 3		Model 4	
Loan purpose	-0.007	(0.124)	-0.010	(0.125)	-0.016	(0.124)	-0.012	(0.124)
Gender	0.040	(0.036)	0.018	(0.030)	0.050	(0.031)	0.032	(0.034)
Age	**0.095***	**(0.050)**	0.028	(0.053)	0.070	(0.058)	0.050	(0.055)
Living city DPI	**0.115*****	**(0.036)**	**0.099*****	**(0.032)**	**0.141*****	**(0.034)**	**0.136*****	**(0.034)**
Monthly income level	**0.124*****	**(0.035)**	**0.112*****	**(0.034)**	**0.100*****	**(0.032)**	**0.090*****	**(0.033)**
Education level	0.019	(0.014)	0.022	(0.016)	**0.037****	**(0.018)**	**0.030***	**(0.016)**
Avg amount of game card	-0.008	(0.012)	-0.020	(0.014)	-0.010	(0.014)	-0.015	(0.014)
ATV shopping durable	0.008	(0.012)	0.015	(0.012)	0.006	(0.012)	0.008	(0.012)
ATV shopping virtual	-0.001	(0.008)	-0.006	(0.006)	-0.003	(0.008)	-0.010	(0.008)
#Outgoing contacts	**-0.019***	**(0.010)**	-0.018	(0.012)	**-0.026****	**(0.012)**	**-0.028****	**(0.012)**
#Office by week	-0.001	(0.012)	0.021	(0.012)	0.020	(0.012)	0.013	(0.012)
#Recreational place by week	-0.033	(0.126)	-0.027	(0.128)	-0.031	(0.132)	-0.027	(0.129)
#Commercial place by week	-0.059	(0.040)	-0.061	(0.042)	-0.050	(0.036)	-0.040	(0.036)
#Public service place by week	0.032	(0.040)	0.011	(0.040)	-0.040	(0.040)	-0.029	(0.040)
IfFinal	**-0.222*****	**(0.089)**	**-0.313*****	**(0.091)**	**-0.344*****	**(0.091)**	**-0.352*****	**(0.090)**
Loan purpose × IfFinal	-0.007	(0.177)	-0.004	(0.167)	-0.006	(0.178)	-0.006	(0.178)
Gender × IfFinal	0.126	(0.110)	**0.185***	**(0.102)**	**-0.199***	**(0.104)**	**-0.178***	**(0.102)**
Age × IfFinal	**0.036***	**(0.020)**	0.030	(0.022)	**0.047****	**(0.025)**	**0.051****	**(0.025)**
Living city DPI × IfFinal	**0.053*****	**(0.014)**	**0.053*****	**(0.014)**	**0.048*****	**(0.014)**	**0.048*****	**(0.014)**
Monthly income level × IfFinal	**0.044*****	**(0.014)**	**0.044*****	**(0.014)**	**0.030****	**(0.015)**	**0.030****	**(0.015)**
Education level × IfFinal	-0.003	(0.019)	**0.030***	**(0.015)**	0.006	(0.017)	**0.027***	**(0.015)**
Avg amount of game card × IfFinal	-0.009	(0.044)	-0.022	(0.043)	- 0.027	(0.043)	-0.016	(0.043)
ATV shopping durable × IfFinal	0.002	(0.002)	0.002	(0.002)	-0.002	(0.002)	-0.001	(0.002)
ATV shopping virtual × IfFinal	-0.007	(0.005)	**-0.011****	**(0.005)**	**-0.010****	**(0.004)**	**-0.016*****	**(0.005)**
#Outgoing contacts × IfFinal	-0.001	**(0.010)**	**-0.027****	**(0.012)**	-0.003	(0.013)	-0.003	(0.013)
#Office by week × IfFinal	0.003	(0.006)	0.010	(0.009)	0.010	(0.008)	0.008	(0.006)
#Recreational place by week × IfFinal	-0.010	(0.115)	-0.005	(0.112)	-0.002	(0.112)	-0.011	(0.114)
#Commercial place by week × IfFinal	**-0.022****	**(0.010)**	**-0.065*****	**(0.011)**	**-0.043*****	**(0.011)**	**-0.050*****	**(0.011)**
#Public service place by week × IfFinal	0.014	(0.048)	0.019	(0.054)	0.016	(0.050)	0.023	(0.055)
Other borrower-related variables	Included		Included		Included		Included	
Other loan-related variables	Included		Included		Included		Included	
Evaluator-related variables	Included		Included		Included		Included	
Log likelihood	-212.48		-238.04		-247.55		-305.16	
#obs.	232		242		268		556	

Table notes are the same as those for Table 6.

D3. Supplementary Analyses and Discussions of Gender Bias

Table D5 Non-default and Approval Rates between Genders across Experimental Groups

Group	Non-default rate		Approval rate		EOR = [(3)/(1)]/[(4)/(2)]
	Female	Male	Female	Male	
	(1)	(2)	(3)	(4)	
1. H & S	87.74%	85.25%	46.29%	44.46%	1.012
2. H & L	89.42%	89.55%	46.76%	47.45%	0.987
3. M & S	90.03%	89.28%	48.11%	46.12%	1.034
4. M & L	93.99%	98.03%	49.11%	42.66%	1.201
5. (H+M) & S & w/o Expl	89.75%	88.18%	48.78%	46.74%	1.025
6. (H+M) & S & w/ Expl	89.70%	89.67%	48.67%	46.80%	1.040
7. (H+M) & L & w/o Expl	93.11%	97.29%	48.29%	42.65%	1.183
8. (H+M) & L & w/ Expl	96.67%	97.55%	48.17%	46.05%	1.056

[a] EOR = Equalized opportunity ratio. The closer to 1 the EOR is, the greater fairness between the genders. The larger the deviation from 1 the EOR shows, the more bias there is toward females (EOR > 1) or males (EOR < 1).

Our findings concerning gender bias are highly generalizable for several reasons. On the one hand, specific sensitive features such as gender are restricted during model training for financial decisions in numerous countries, a measure taken to ensure fairness and equality. Unfortunately, recent studies have revealed that advanced machine-learning models, such as XGBoost, often exhibit a *triangulation* effect. This effect involves effectively combining observed permissible features (e.g., non-sensitive demographics) to infer the impact of unobserved restricted sensitive features on the outcome. The consequence is heightened inequality and diminished borrower welfare, as demonstrated by various studies (Fuster et al. 2022, Lu et al. 2023a, Hu et al. 2023). Despite not incorporating gender as a feature for machine-learning model training and human evaluators' assessments, we posit that machines can still introduce biases based on gender due to the triangulation effect.

On the other hand, whereas our study illustrates that humans can mitigate gender bias, in the presence of large information volumes and machine explanations, our identified mechanism did not find any indication that humans intentionally sought to address gender bias. Put differently, the reduction of gender bias observed after involving human evaluators in Group 8 is an indirect outcome resulting from humans rethinking and associating relevant features. While such benefits may be context-dependent, we assert that our findings have broader applicability in terms of managerial implications. This is because we reveal that humans can leverage additional information to rectify machine errors, rather than being confined to existing knowledge or biases.

Furthermore, we recognized the body of literature investigating instances of human-induced discrimination across various scenarios. The results are overall context-dependent. To illustrate, differing from our findings, Bartlett et al. (2022) demonstrated that lenders tend to use proxies for unobservable factors, resulting in biased decisions against minority borrowers. Furthermore, Chen et al. (2017) revealed that in peer-to-peer lending markets, lenders might disfavor female borrowers. In such situations, algorithms or machines possess the potential to mitigate biases and enhance revenue outcomes. This represents an instance where human-AI collaboration can yield a synergistic effect, surpassing the sum of its parts. The efficacy of the solution relies more on the algorithm itself. Crucially, as discussed in previous paragraphs, AI or machines might also cause decision biases in many contexts. Our study considered this type of situation and introduced a novel perspective, aiming to guide humans in improving collaboration with AI or machines. In this context, our emphasis is on exploring how humans can play a pivotal role in mitigating gender biases generated by machines. To this extent, our findings do not go against the previous studies in that we pave the way toward the realization of "1+1>2." On a broader scale, our discoveries can be applied to a more extensive range of scenarios by investigating how humans can act as curators, rectifying errors made by either machines or other humans.