informs
https://pubsonline.informs.org/journal/isre

INFORMATION SYSTEMS RESEARCH
Articles in Advance, pp. 1–25
ISSN 1047-7047 (print), ISSN 1526-5536 (online)

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

Tian Lu,[a] **Yingjie Zhang**[b,*]

[a] Department of Information Systems, W. P. Carey School of Business, Arizona State University, Tempe, Arizona 85287; [b] Guanghua School of Management, Peking University, Beijing 100871, China

*Corresponding author

Contact: lutian@asu.edu, https://orcid.org/0000-0003-3730-1897 (TL); yingjiezhang@gsm.pku.edu.cn, https://orcid.org/0000-0003-0745-2563 (YZ)

Received: May 21, 2023
Revised: January 6, 2024; March 13, 2024
Accepted: April 2, 2024
Published Online in Articles in Advance: May 2, 2024

https://doi.org/10.1287/isre.2023.0305

Abstract. With the explosive growth of data and the rapid rise of artificial intelligence and automated working processes, humans inevitably fall into increasingly close collaboration with machines as either 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 with human performance, it is essential to explore whether human–machine collaboration is valuable and, if so, why. Recent studies propose 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 these 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 propose 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 reveal that the coexistence 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 demonstrate that humans can 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 nontrivial implications that are both theoretical and practical.

History: Param Vir Singh, Senior Editor; Tianshu Sun, Associate Editor.
Funding: This work was supported by the National Natural Science Foundation of China [Grant 72272003].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/isre.2023.0305.

Keywords: decision making • gender biases • human–machine collaboration • information processing • machine explanations • microfinance • 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 as both 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 finds 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 because of distrust of machines (Jacovi et al. 2021) or overreliance 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 overcautious (Lu et al. 2023b) or hyper-focused on details (Wang et al. 2023d).

To address the urgent, essential question regarding how to efficiently change humans' responses to machines from either aversion or overreliance 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 predominantly concentrate on technical solutions and lack 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 typically conduct laboratory experiments or simulations alone. This approach proves challenging as participants tend to present differently and participate more actively in a controlled laboratory 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 as follow: (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 focus 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 these 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, whereas machines explored more possibilities using other sources of information, including shopping and off-line 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 noninformative.

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 intergender 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 multifold. First, it adds to the emerging literature on human–machine collaboration. Whereas a few of the most recent studies investigate whether humans and machines complement each other in decision making in different contexts (e.g., Luo et al. 2019, Cao et al. 2021, Zhang et al. 2024), the majority suggest 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 for 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 largely suggest that displaying (feature-based) machine explanations 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 propose the utilization of multisource data to alleviate algorithmic discrimination and sample biases. In fact, there is evidence that alternative data sources eliminate biases related to race and socioeconomic factors (Lu et al. 2023a). However, machine failure is already 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 and 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 explore the pros and cons of human–machine collaboration in decision making. For example, studies show that most statistical models exceed or approach the judgment accuracy of the average clinician (Camerer 2019). Machine algorithms are extensively shown to manage substantial amounts of data more proficiently than humans (Peukert et al. 2023, Wang et al. 2023d). 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 show the superiority of human–machine collaborations over both full machine automation and human-only operations (Fügener et al. 2022) and 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. 2023c, Zhang et al. 2024) 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) show 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 also reveals 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 because of 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 outperform humans in many jobs, especially low-skilled, repetitive, and dangerous ones (Autor and Dorn 2013). Conversely, Fügener et al. (2021) warn that we must also attend to humans' overreliance 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 examines 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. 2020, 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. 2021). However, this approach should be employed with caution. Schmidt et al. (2020) indicate that offering unintuitive explanations (i.e., those dealing with features with which humans are unfamiliar) may fail to boost humans' trust in machines. Rudin (2019) also cautions 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, whereas some previous studies explore the impact of machine explanations on human–machine collaboration, few delve 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 reveals that humans can dynamically adjust the importance they attribute to available information and adapt their mental models based on machine explanations. Additionally, their findings highlight 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, whereas Bauer et al. (2023) only attend 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) may have been influenced by their use of online laboratory experiments. By their nature, laboratory 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 laboratory 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 microfinance context to examine individuals' decision making in a more natural setting.

2.3. Investors' Decision Making in Microfinance

Many scholars focus on individual investors' decision making in microfinance businesses, including peer-to-peer (P2P) lending, crowdfunding, and microloans. A subset of the literature reveals 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 also identify biases in microfinance investors' decisions, including preferences regarding gender (Chen et al. 2017) or location (Lin and Viswanathan 2016). Recent research pays attention to the value of machine-assisted tools in financial decision making. For example, Ge et al. (2021) find 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 2022), 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 microfinance 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 overreliance 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 coworking 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 quickly, 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 2 information processing and reasoning. To encourage individuals to embrace system 2 processing, certain conditions must be met. Specifically, because 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 (Sun and Taylor 2020, Bauer et al. 2023). Specifically, new attributes provide novel pieces of information that enhance one's recognition of the decision tasks and domain (He et al. 2021). 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: (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. 2021). 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. 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 more than 250,000 borrowers by 2018 with unsecured microloans of approximately US$465. The company uses only the owner's money for lending, and the 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 one to seven 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 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 focus only on loans with a term of one, two, or three 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 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, maintaining the approval rate specified.

3.2. Experimental Setup

3.2.1. Implementation of Treatment I: Information Complexity.

Inspired by the dual-process theories of reasoning, we introduce two factors that could influence human evaluators' decision making in collaboration with machines in Section 2.4. As the first step, we utilize 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 suggest 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. Online Table A.1 describes these features.

3.2.2. Machine Preparations and Implementation of Treatment II: Machine Explanations. Because 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 comprise borrowers who submitted loan applications June 1–30, 2017. For these sampled borrowers, the human evaluators 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. Because the loan term was no longer than three months, a five-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 this 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 hyperparameter 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, multilevel 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 seven-day training period, all human evaluators reached a stable evaluation performance. Please refer to Online Appendix A.2 for detailed information on the human evaluators and the training procedure.

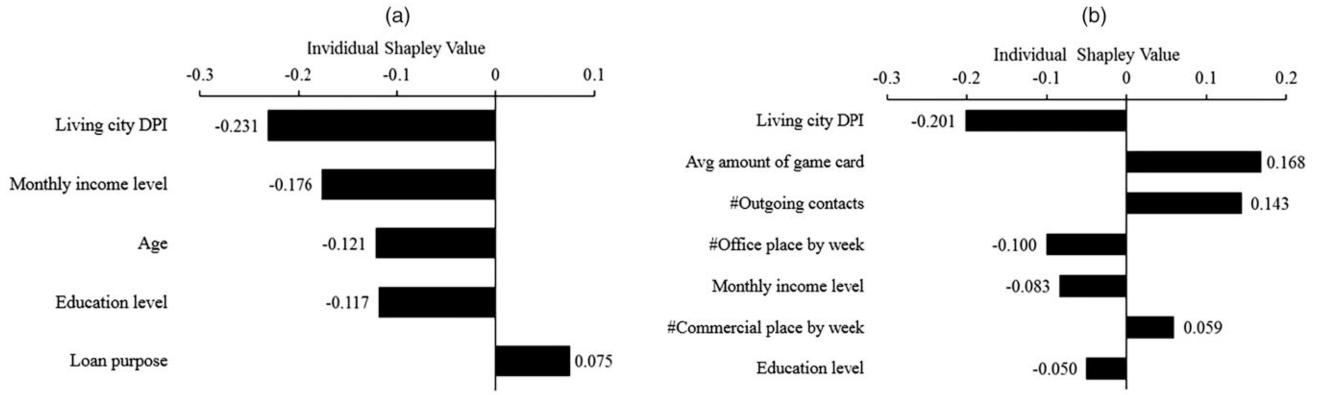
With the pretrained prediction models, we were able to design the second treatment. Specifically, to prepare the machine explanation information based on the machine learning algorithms, we implemented a SHAP analysis method, which yields Shapley values representing the average expected marginal contribution 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.

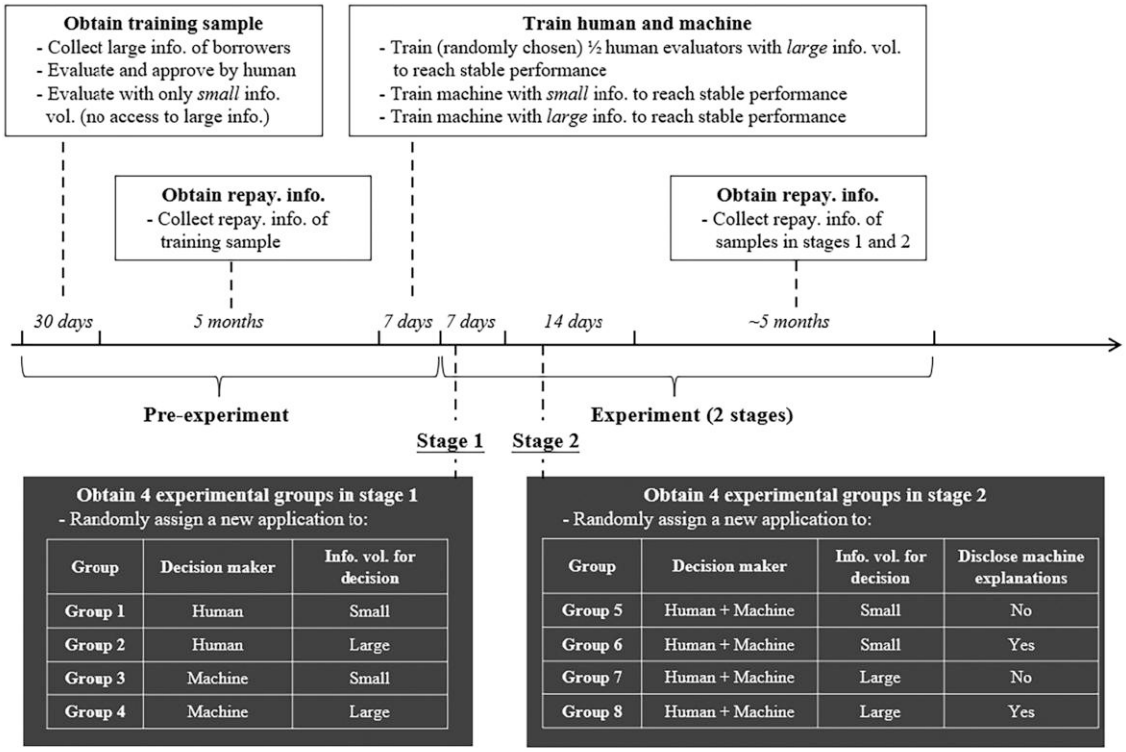
3.3.1. Experimental Stage 1. The first stage began on December 8, 2017, and lasted for one week. The relatively short term of the treatments helped tease out the

Figure 1. Important Features in Machines' Decision-Making Processes



Notes. (a) Decisions with small information volume. (b) Decisions with large information volume. These features rank in the top five or seven 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 2. Experimental Process


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:
Obtain 4 experimental groups in stage 2
- 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

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

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 pretrained 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.

3.3.2. 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 Online Figure A.2, 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 Online Figure A.2a). 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 nondefaulters (see Online Figure A.2b). 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 because of the 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 five months.

3.4. Experimental Data

We obtained our experimental data after completing repayment information collection. The data set 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).

There was a total of 23,805 loans in the eight 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 educational background. On average, the human evaluators had been working for the company for approximately one year, and those with high, medium, and low levels of historical decision performance were evenly distributed between the groups (i.e., around one third 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 intergroup 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 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 Online Figure B.1 confirm the consistency.

Noticing these diverse patterns, we then further decomposed the decision-making behavior of human evaluators after they had observed the machines' suggestions. Specifically, in Figure 4, we compare the decision consistency between humans (initial decision) and machines (in Figure 4(a)) and calculate the adjustment

Table 1. Randomization Check

		Loan characteristics			Borrower characteristics				
Group	Number of observations	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	Number of 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

Notes. H = human decision, M = machine decision, H + M = human + machine decision, w/o Expl = without AI explanations, w/ Expl = with AI explanations. Loan purpose: 1 = consumption, 0 = others (e.g., for emergency). Gender: 1 = female, 0 = male. 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. Education level: 1 = middle school or below, 2 = vocational school, 3 = high school, 4 = technical school, 5 = undergraduate, 6 = graduate or above. Evaluator months working: 1 = not longer than 6 months, 2 = 6–12 months, 3 = 13–18 months, 4 = longer than 18 months. Evaluator historical (decision) accuracy: 1 = low (default rate >15%), 2 = medium (10% < default rate < 15%), 3 = high (default rate <10%). Refer to Table A2 in Online Appendix A.1 for descriptive statistics on evaluator historical accuracy. 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). For every feature, the values show no significant differences across the groups based on the *F*-test.

ratios when inconsistency arose (in Figure 4(b)). 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 4(a), the agreement proportion was smaller with the large information volume (83.78% consistency in group 5

Figure 3. Default Rates of Experimental Groups

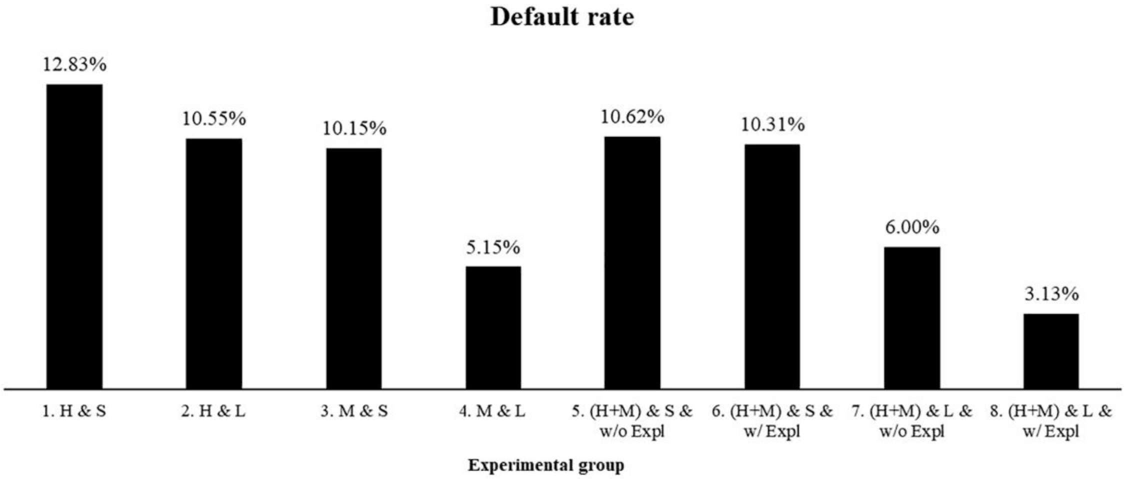


Table 2. Comparison of Default Rates Among Different Experimental Groups

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***

Notes. As our experiment comprises multiple treatments, we follow 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.

$^*p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

versus 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 versus 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 versus group 8, p-value = 0.026).

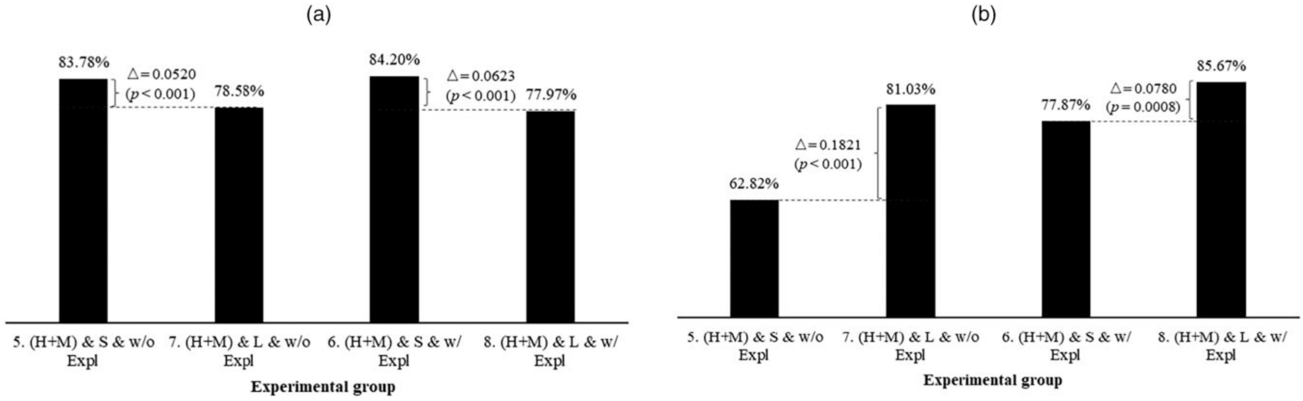
5. Mechanism Examinations

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

Figure 4. Consistency and Following Between Humans and Machines



Notes. (a) Ratio of decision consistency between humans and machines. (b) Ratio of following machines' decisions (conditional on decision inconsistency between humans and machines). p-values are multiplicity-adjusted values (List et al. 2019) based on between-group t tests.

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 machines' decision-making process. Features with significant coefficients in the regressions were important features.

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 volume 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]

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 disposable personal income (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

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)	
Dependent variable: 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	
Number of observations	8,804		17,531	

Notes. 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. 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. Standard errors are in parentheses. Significant results are in bold.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

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)	
Dependent variable: `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	
Number of observations	8,944		17,798	

Notes. 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. 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. Standard errors are in parentheses. Significant results are in bold.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

contacts. In contrast, the machines explored additional sources 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 off-line 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 are confirmed to have predictive advantages using novel features from alternative data sources (Zhou et al. 2021, Lu et al. 2023a). 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 disappears if either of the two conditions is removed. In order to understand the human

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

Note. Correlations are based on all loan samples. Relatively large values are in bold.

evaluators' behavior, we conducted regression tests using observations in which the human evaluators' initial decisions differed from those of the machines.

We employ probit models in which the DV was `IfApprove` and the independent variables include all available loan features. To understand how machines' recommendations influence humans' decision-making processes, we compared the discrepancies between human evaluators' initial and final decisions. Empirically, we define a new binary indicator, `IfFinal`, which equals one if the approved decision was made after a machine recommendation was present. Again, we include 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 report the results in Tables 6 and 7. With small information volumes (groups 5 and 6 in Table 6) and large information volumes but no machine explanations (group 7 in Table 7), the factors that explain the human evaluators' final approval decisions remain similar to those in the first stage (shown in Tables 3 and 4).

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

	Group 5 (humans' final decision)		Group 5 (initial vs. final)		Group 6 (humans' final decision)		Group 6 (initial vs. final)	
Dependent variable: `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	
Number of observations	468		936		461		922	

Notes. Models 1–4 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. 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. Standard errors are in parentheses. Significant results are in bold.

*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

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)	
Dependent variable: `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)	**−0.154****	**(0.069)**	0.032	(0.034)
Age	**0.130****	**(0.056)**	0.073	(0.059)	**0.100*****	**(0.018)**	0.051	(0.057)
Living city DPI	**0.135*****	**(0.025)**	**0.098*****	**(0.024)**	**0.188*****	**(0.030)**	**0.140*****	**(0.033)**
Monthly income level	**0.076****	**(0.032)**	**0.032*****	**(0.010)**	**0.140*****	**(0.034)**	**0.110*****	**(0.033)**
Education level	**0.191****	**(0.081)**	**0.130*****	**(0.040)**	**0.032***	**(0.019)**	**0.030***	**(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)	**−0.020*****	**(0.005)**	−0.010	(0.008)
#Outgoing contacts	**−0.036*****	**(0.010)**	**−0.015***	**(0.008)**	**−0.025****	**(0.012)**	**−0.022***	**(0.012)**
#Office by week	**0.028****	**(0.011)**	0.067	(0.042)	**0.027****	**(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	**−0.044***	**(0.025)**	−0.022	(0.024)	**−0.111*****	**(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			**−0.337*****	**(0.091)**			**−0.349*****	**(0.091)**
Loan purpose × *IfFinal*			−0.170	(0.164)			−0.006	(0.178)
Gender × *IfFinal*			0.149	(0.134)			**−0.185***	**(0.105)**
Age × *IfFinal*			**0.056****	**(0.028)**			**0.048****	**(0.025)**
Living city DPI × *IfFinal*			**0.053*****	**(0.014)**			**0.048*****	**(0.014)**
Monthly income level × *IfFinal*			**0.044*****	**(0.014)**			**0.030****	**(0.015)**
Education level × *IfFinal*			**0.061*****	**(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)			**−0.010****	**(0.004)**
#Outgoing contacts × *IfFinal*			**−0.021****	**(0.010)**			−0.003	(0.013)
#Office by week × *IfFinal*			**0.017***	**(0.010)**			0.008	(0.006)
#Recreational place by week × *IfFinal*			−0.095	(0.108)			−0.007	(0.114)
#Commercial place by week × *IfFinal*			**−0.022***	**(0.014)**			**−0.052*****	**(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	
Number of observations	638		1,276		649		1,298	

Notes. Models 1–4 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. 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. Standard errors are in parentheses. Significant results are in bold.

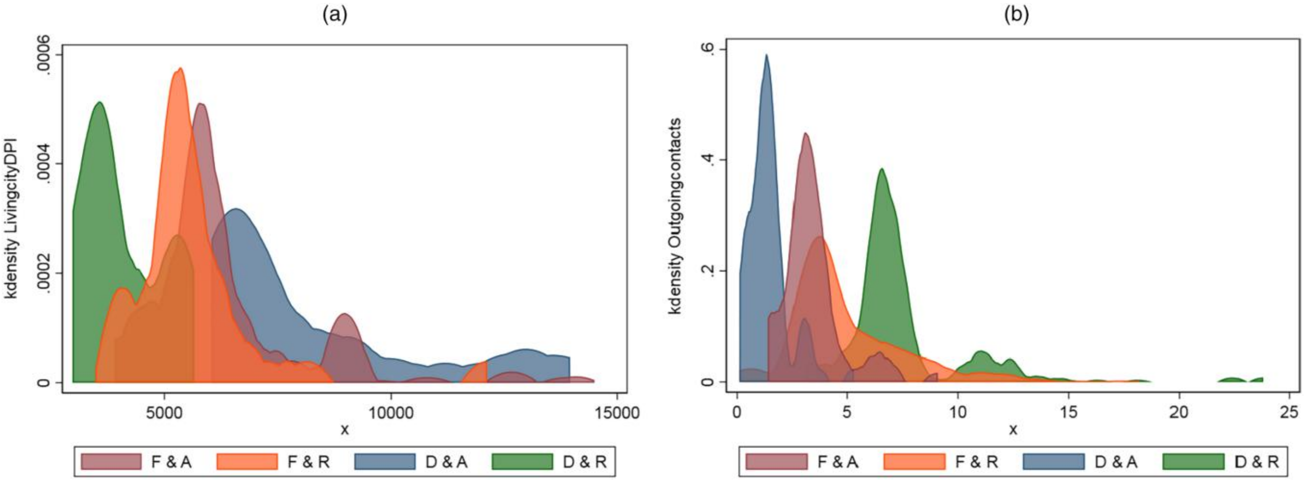
*$p < 0.10$; **$p < 0.05$; ***$p < 0.01$.

For example, with the small information volume, the interaction terms for three features (i.e., applicant age, living city DPI, and monthly income level) present 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 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 Online Appendix C.2. 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 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 tend to stick with their initial opinions. Considering that the machines incorporate 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. To be specific, we visualize the distributions

Figure 5. (Color online) Feature Distributions of Diverse Cases (Group 7)



Notes. (a) Distributions of living city DPI. (b) Distributions of #outgoing contacts. All distributions are based on the samples in which human evaluators' initial decisions were inconsistent with machines' decisions. 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.

using four subsamples, which are separated by two standards: whether human evaluators ultimately accepted or rejected the applications (A versus R) and whether humans followed or continued to disagree with the machines' recommendations (F versus D). Interestingly, we observe 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 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 Online Figure C.2.

This result, however, was not found in group 8. Interestingly, in Models 3 and 4 of Table 4, we notice 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 explain 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 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 observe that, with the presence of large information volumes and machine explanations, humans reconsidered an interesting feature, ATV shopping virtual. This feature was not 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 because of its correlations with other features. We conjecture that the attention to the ATV shopping virtual feature stems 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 with which the evaluators were unfamiliar (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 zero (refer to Online Figure A.1(a); the median is zero). Such a distribution leads to human evaluators perceiving those features as noninformative. The literature suggests that humans are good at building connections between given information and other relevant, familiar, or understandable information in cognitive processing (Hollnagel 1987, Bråten and Samuelstuen 2007). Because game cards are typical virtual goods and ATV shopping virtual had many more salient nonzero values (Online Figure A.1(b); the median is 8.70), human evaluators are likely to attend more to this feature when making decisions.

In Online Appendix C.4, we compare 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 because of the machines' approval recommendations) and those saved by human evaluators. We show 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 in which 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 demonstrate in Section 6.2). Moreover, we conducted a straightforward post hoc analysis in Online Appendix C.4 to clarify the allocations of different loan types by humans, machines, and collaborative efforts. This provides additional insights into how machines and humans can 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 take responsibility for handling borderline cases, and humans have the potential to invoke active rethinking to correct machines' mistakes in the random cases (i.e., 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 show 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 decompose the heterogeneity regarding individual evaluators' characteristics. 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 (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, availability of machine explanations, and experience levels. We also include all lower level interaction terms in the regression. We present the estimated coefficients in Table 8, wherein Model 1 considers the default rate as DV and includes humans' independent decisions only, whereas Models 2–4 are in the human–machine collaboration modes. Specifically, we replicate our mechanism tests with heterogeneous experience levels: whether a loan defaulted is Model 2's DV, whether initial decisions were consistent is Model 3's DV, and whether to follow machines' decisions is Model 4's DV. Note that the estimation of Model 4 incorporates only samples for which human evaluators' initial decisions were inconsistent with machine decisions. Additionally, we offer more comprehensive heterogeneity analyses with alternative characteristics in Online Appendix D.1.

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 observe 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 focus 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 Online Appendix D.2, 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 focus on the final performance as measured by the nondefault rates. We recorded the statistics of each group in Online Table D.5. We observe that, with large information volumes (i.e., group 4), machines tended to favor female applicants because the nondefault 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% versus 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 G be the gender indicator ($G = 0$ or 1) and $Y = 1$ and $\hat{Y} = 1$ be the correct and actual positive outcomes (i.e., a loan application being approved in our context), respectively. "Correct" here means that nondefault loans (observed from the repayment performance of the approved loans) got

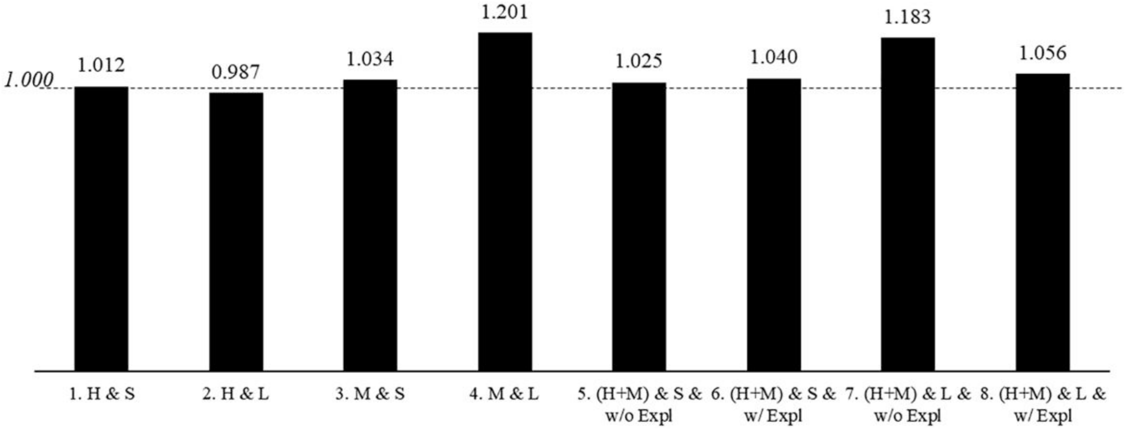
Table 8. Heterogeneity Analysis of Human Evaluators' Months Working (Probit Model)

	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	
Number of observations	2,716		11,727		5,603		2,216	

Notes. 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 Online Appendix C.2. 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. Standard errors are in parentheses. Significant results are in bold.

$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure 6. Equalized Opportunity Ratio on Gender



Note. Refer to Online Table D.5 for complete values.

approved. As such, equalized opportunity means $Pr(\hat{Y}=1 \mid G=0, \quad Y=1)=Pr(\hat{Y}=1 \mid G=1, \quad Y=1)$. Applying this criterion to our context, EOR describes the decision biases between genders as follows: $EOR=\frac{Appr(G=0)/NonD(G=0)}{Appr(G=1)/NonD(G=1)}$, where $Appr(G=0)$ and $Appr(G=1)$ refer to the approval rates for females and males, and $NonD(G=0)$ and $NonD(G=1)$ refer to the nondefault rates for females and males (calculated within female or male groups), respectively. The closer EOR is to one, the greater the fairness is between the genders. The larger the deviation from one, 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 learn 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 mainly due 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 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 can 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 data set 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 provide a comprehensive discussion about how our findings concerning gender biases can be extrapolated to other contexts in Online Appendix D.3.

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 dove 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 apply 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 show 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 demonstrate that humans tend to utilize what they specialize in (i.e., small information volumes, group 1 versus 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 versus 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 is 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 versus group 8). This is owed to the fact that machines' superiority in tackling prohibitively (for humans) complex tasks to achieve satisfactory predictions is constrained by information availability (also refer to the comparison between group 3 versus group 4 in Figure 3). On top of limited information, humans cannot become smarter than machines. Notably, we notice that a few recent studies focus on the value of machine explanations to human–machine collaboration (e.g., Jacobs et al. 2021, Bauer et al. 2023). However, our study suggests that contingent factors, such as task complexity, impact the effect of machine explanations. Although machine explanations provide humans with more reference information, humans may not take advantage of them because of insufficient motivation to deeply involve themselves in decision making (Speier 2006). Instead, humans are 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 collaborations 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 nontrivial 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 are 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 create additional value. If only small data are 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 in which 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 in which 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). Finally, 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 in which 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, whereas 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 are guided by theory and offer 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 delve into the difficulties faced by humans when attempting to articulate or summarize the rules guiding their decision-making processes (Hu et al. 2022). Compared with 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 emphasizes 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. Recognizing that two-stage designs may be practically infeasible, we suggest the potential relevance of our findings in scenarios in which 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 overreliance on machines or foster distrust because of 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 with 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 focuses 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 considers 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.

Acknowledgments

Both authors contributed equally and share the joint first authors. The authors thank the senior editor, the associate editor, and the reviewers for their insightful comments and constructive suggestions throughout the review process.

Endnotes

[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.

[3] The groups with large information volumes had access to multisourced 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.

[4] In Online Appendix C.1, we conduct multiple robustness checks. First, we reran our regressions within each experimental group using different samples. The results indicate that the human evaluators' initial decisions did not involve any learning from the machines' recommendations. This also confirms that the comparisons between the two stages in our experiment are reasonable. As another robustness check, we employed decision tree approaches to infer the decision rules implemented by human evaluators. The results in Online Figure C.1 confirm the consistency. Additionally, we incorporate an alternative DV to offer more insights into how humans and machines reached the same or different initial decisions.

[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 alleviate this concern in Online Appendix C.3.

References

Alibaba Cloud (2018) 6 fields where artificial intelligence are surpassing human. Accessed January 30, 2024, 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. *Organ. Sci.* 33(1):149–169.

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

Autor DH, Dorn D (2013) How technology wrecks the middle class. *New York Times* 24(2013):1279–1333.

Bauer K, von Zahn M, Hinz O (2023) Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing. *Inform. Systems Res.* 34(4):1582–1602.

Becker HS (1995) The power of inertia. *Qualitative Sociol.* 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 Learn.* 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. *Psych. Bull.* 119(2):197–253.

Camerer CF (2019) Artificial intelligence and behavioral economics. Agrawal A, Gans J, Goldfarb A, eds. *The Economics of Artificial Intelligence: An Agenda* (University of Chicago Press, Chicago), 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, Cambridge, MA.

Chapman LJ, Chapman JP (1967) Genesis of popular but erroneous psychodiagnostic observations. *J. Abnormal Psych.* 72(3):193–204.

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 Res.* 17(4):553–583.

Chen V, Liao QV, Vaughan JW, Bansal G (2023) Understanding the role of human intuition on reliance in human-AI decision-making with explanations. Preprint, submitted January 18, https://arxiv.org/abs/2301.07255.

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

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

Commerford BP, Dennis SA, Joe JR, Ulla JW (2022) Man vs. machine: Complex estimates and auditor reliance on artificial intelligence. *J. Accounting Res.* 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. *Inform. Systems Res.* 23(4): 1093–1109.

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

de Véricourt F, Gurkan H (2023) Is your machine better than you? You may never know. *Management Sci.*, ePub ahead of print May 25, https://doi.org/10.1287/mnsc.2023.4791.

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

Endsley MR (1995) Toward a theory of situation awareness in dynamic systems. *Human Factors* 37(1):85–104.

Evans JSBT (2003) In two minds: Dual-process accounts of reasoning. *Trends Cognitive Sci.* 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. *Inform. Systems Res.* 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 Inform. Systems Quart.* 45(3b):1527–1556.

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

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

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

Germann M, Merkle C (2022) Algorithm aversion in financial investing. Preprint, submitted November 6, https://dx.doi.org/10.2139/ssrn.3364850.

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

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical vs. mechanical prediction: A meta-analysis. *Psych. Assessment* 12(1):19–30.

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 (2021) Enhancing user privacy through ephemeral sharing design: Experimental evidence from online dating. Preprint, submitted January 24, https://dx.doi.org/10.2139/ssrn.3740782.

Hollnagel E (1987) Information and reasoning in intelligent decision support systems. *Internat. J. Man Machine Stud.* 27(5–6):665–678.

Hu X, Huang Y, Li B, Lu T (2022) Uncovering the source of machine bias. Preprint, submitted January 9, https://arxiv.org/abs/2201.03092.

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

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

Icard TF (2018) Bayes, bounds, and rational analysis. *Philos. Sci.* 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. *Proc. 2021 ACM Conf. Fairness Accountability Transparency* (Association for Computing Machinery, New York), 624–635.

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

Kahneman D (2011) *Thinking, Fast and Slow* (MacMillan, New York).

Keil M, Tan BCY, Wei KK, Saarinen T, Tuunainen V, Wassenaar A (2000) A cross-cultural study on escalation of commitment behavior in software projects. *Management Inform. Systems Quart.* 24(2): 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. Preprint, submitted February 3, https://arxiv.org/abs/2202.01602.

Kunimoto C, Miller J, Pashler H (2001) Confidence and accuracy of near-threshold discrimination responses. *Consciousness 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. *Organ. Behav. 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 Sci.* 62(5):1393–1414.

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

Liu M, Tang X, Xia S, Zhang S, Zhu Y, Meng Q (2023) Algorithm aversion: Evidence from ridesharing drivers. *Management Sci.*, ePub ahead of print October 3, https://doi.org/10.1287/mnsc.2022.02475.

Lou B, Wu L (2021) AI on drugs: Can artificial intelligence accelerate drug development? Evidence from a large-scale examination of bio-pharma firms. *Management Inform. Systems Quart.* 45(3): 1451–1482.

Loutfi E (2019) *What Does the Future Hold for AI-Enabled Coaching* (Chief Learning Officer, New York).

Lu SF, Rui H, Seidmann A (2018) Does technology substitute for nurses? Staffing decisions in nursing homes. *Management Sci.* 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. *Management Inform. Systems Quart.* 47(4):1517–1556.

Lu X, Huang Y, Zhang Y, Shen L (2023b) At the cost of active thinking: Investigating the role of presentation explicitness in human-AI collaboration. Preprint, submitted August 27, https://dx.doi.org/10.2139/ssrn.4547893.

Lu J, Lee D, Kim TW, Danks D (2020) Good explanation for algorithmic transparency. Preprint, submitted January 7, https://dx.doi.org/10.2139/ssrn.3503603.

Luo X, Qin MS, Fang Z, Qu Z (2021) Artificial intelligence coaches for sales agents: Caveats and solutions. *J. 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 Sci.* 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. *J. Consumer Res.* 25(4):335–352.

Marcotte DE (1998) The wage premium for job seniority during the 1980s and early 1990s. *Indust. Relations* 37(4):419–439.

Mohseni S, Yang F, Pentyala S, Du M, Liu Y, Lupfer N, Hu X, Ji S, Ragan E (2021) Machine learning explanations to prevent overtrust in fake news detection. *Proc. Internat. AAAI Conf. Web Social Media*, vol. 15 (AAAI Press, Palo Alto, CA), 421–431.

Oskamp S (1965) Overconfidence in case-study judgments. *J. Consulting Psych.* 29(3):261–265.

Peukert C, Sen A, Claussen J (2023) The editor and the algorithm: Recommendation technology in online news. *Management Sci.*, ePub ahead of print October 17, https://doi.org/10.1287/mnsc.2023.4954.

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

Rai A (2020) Explainable AI: From black box to glass box. *J. Acad. Marketing Sci.* 48(1):137–141.

Roth AE (1988) Introduction to the Shapley value. *The Shapley Value* (Cambridge University Press, Cambridge, UK), 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. *Psych. Bull.* 66(3):178–200.

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

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

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. *Cutter Bus. Tech. J.* 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* (Edward Elgar Publishing, Cheltenham, UK), 3–14.

Song QC, Tang C, Wee S (2021) Making sense of model generalizability: A tutorial on cross-validation in R and shiny. *Adv. Methods Practices Psych. Sci.* 4(1).

Speier C (2006) The influence of information presentation formats on complex task decision-making performance. *Internat. J. Human Comput. Stud.* 64(11):1115–1131.

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

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 Sci.* 68(2):846–865.

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

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 Sci.*, ePub ahead of print November 14, https://doi.org/10.1287/mnsc.2022.03518.

Teodorescu MHM, Morse L, Awwad Y, Kane GC (2021) Failures of fairness in automation require a deeper understanding of human-ML augmentation. *Management Inform. Systems Quart.* 45(3):1483–1500.

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

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

Wang W, Benbasat I (2016) Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents. *J. Management Inform. 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 Sci.*, ePub ahead of print October 11, https://doi.org/10.1287/mnsc.2021.00588.

Wang W, Yang M, Sun T (2023c) Human-AI co-creation in product ideation: The dual view of quality and diversity. Preprint, submitted December 20, https://dx.doi.org/10.2139/ssrn.4668241.

Wang W, Liu X, Zhang X, Hong Y (2023d) Knowledge trap: Human experts distracted by details when teaming with AI. Preprint, submitted April 3, https://dx.doi.org/10.2139/ssrn.4395858.

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 Res. Appl.* 36:100873.

Wang L, He Y, Huang N, De Liu XG, Chen G (2023a) The role of AI assistants in livestream selling: Evidence from a randomized field experiment. Preprint, submitted February 25, https://dx.doi.org/10.2139/ssrn.4365103.

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

Zhang M, Sun T, Luo L, Golden J (2024) Consumer and AI co-creation: When and why nudging human participation improves AI creation. *USC Marshall School of Business Research Paper Sponsored by iORB, No.* Forthcoming.

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.