How Forced Intervention Facilitates AI Adoption

Xinyu Cao[1], Chenshan Hu[2], Jiankun Sun[3], Dennis J. Zhang[4]
1. CUHK Business School, The Chinese University of Hong Kong
2. Leeds School of Business, University of Colorado Boulder
3. Imperial Business School, Imperial College London
4. Olin Business School, Washington University in St. Louis

Problem Definition: While artificial intelligence (AI) technologies increasingly become powerful and useful in operations, human workers often resist adopting algorithms, known as algorithm aversion. This aversion can undermine the algorithm performance in practice. While numerous studies explored short-term mitigation strategies for such aversion, this paper investigates whether and why forced interventions can promote AI adoption and reduce algorithm aversion in practice.

Methodology/Results: Data from a leading online education company reveal that sales workers underutilize a new matching algorithm and often selectively use it on low-quality leads. The company conducted a field experiment where sales workers were forced to use or not use the algorithm for three weeks. Experimental results show that forcing workers to use the algorithm during the experiment causally increases their algorithm usage over the month after the experiment by 15.8 percentage points. We develop a theoretical model to derive empirical strategies for exploring the mechanisms behind this improvement. Contrary to the traditional literature focusing on habit formation, our findings suggest learning is a key driver for algorithm adoption among workers over the month after the experiment. Specifically, forced algorithm use allows workers to experience the unbiased algorithm performance and positively adjust their beliefs about it. Consequently, after the experiment, workers use the algorithm not only more frequently but also more on high-quality leads.

Managerial Implications: The study empirically shows that forced intervention can effectively improve persistent algorithm use after the intervention, which is crucial for continuous development of the algorithms. More importantly, forced intervention breaks the vicious cycle of biased beliefs and selective usage by enabling workers to form unbiased evaluation of the algorithm efficacy and mitigate selective adoption on low-quality cases. This suggests that firms can implement extrinsic interventions or educational programs to help workers recognize the benefits of algorithms and develop unbiased beliefs about their capabilities, thus facilitating sustained algorithm usage.

Key words: AI adoption, algorithm aversion, learning, habit formation, field experiment

1. Introduction

With the explosion of data, the improvement of computing power, and the rapid development of machine learning algorithms, a large number of companies start to utilize artificial intelligence (AI) technologies or smart algorithms to improve their business practices. According to a survey by PwC in 2023, 73% of US companies have already adopted AI in their business, and 54% of the companies surveyed have implemented generative AI in some business areas.[1] In particular, AI technologies are vastly transforming operations of numerous companies across various functions, such as product and service development,

[1] https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-predictions.html

supply chain management, and customer service. A majority of these companies have witnessed revenue growth in the business functions using AI, as reported in a survey by McKinsey & Company.[2]

In practice, the successful application of AI hinges on the broad adoption and sustained usage by users, particularly in its early development stage, since AI improvement relies on user-generated data. Unlike traditional technology innovations that are usually adopted after reaching maturity and demonstrating superior performance, AI is often implemented at an early stage before it can consistently outperform humans. In fact, AI may even cause short-term losses initially before delivering long-term gains, following a "J-curve" trajectory in performance over time (Vaccaro et al. 2024, McElheran et al. 2025). Early implementation of AI is often important in that the improvement of AI performance follows an iterative and data-driven process with continuous learning and refinement (Domingos 2012, Hospedales et al. 2021). The key driver for this process is the well-recognized virtuous feedback loop among user adoption, data generation, and AI improvement—termed *data network effects*: More users of AI generate more usage data for algorithms to learn from, which enables algorithm performance improvement and in turn attracts more users (Gregory et al. 2021, Haftor and Costa-Climent 2023, Levine and Jain 2023). For example, Tesla's Autopilot technology is continuously learning and improving using the data from customers driving and using Autopilot. As explained by Elon Musk, the CEO of Tesla, each driver using the autopilot system essentially becomes an "expert trainer for how the autopilot should work."[3] In particular, a common challenge in the initial phase of AI implementation is the "cold-start" problem: The algorithm does not have enough data to achieve superior performance (e.g., Schein et al. 2002, Ye et al. 2023). Therefore, encouraging algorithm adoption and sustained usage during its cold-start stage—even when the algorithm performance is not yet superior—is generally crucial for the algorithm improvement and operational performance gains.

However, it has been documented that humans tend to be averse to use algorithms, a phenomenon known as algorithm aversion (e.g., Dietvorst et al. 2015, Dietvorst and Bharti 2020). This aversion can be especially pronounced during the early stages of algorithm development, when the algorithm performance may not yet significantly exceed human capabilities. Numerous studies in operations management have explored in various contexts humans' aversion to use algorithms (e.g., Knight et al. 2022, Lin et al. 2023) or their biases against algorithmic advice (e.g., Sun et al. 2022, Caro and de Tejada Cuenca 2023, Liang et al. 2023). These studies have proposed various short-term solutions to mitigate human biases against algorithms, which typically already exhibit superior performance. However, it remains unclear whether these solutions can foster sustained algorithm adoption among workers throughout the cold-start stage of an algorithm, when the algorithm performance is not yet superior but facilitating its adoption is crucial for continued improvement of the algorithm performance. Therefore, it is both theoretically and

[2] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-AIs-breakout-year/

[3] https://fortune.com/2015/10/16/how-tesla-autopilot-learns/

practically important to explore methods to reduce algorithm aversion and enhance persistent algorithm usage, particularly during the early stages of algorithm development.

To address this problem, we fall back to the behavioral literature to explore how short-term extrinsic interventions can shape persistent behavior. For example, some studies have explored the impacts of financial incentives on various behaviors after the intervention, such as exercising (e.g., Charness and Gneezy 2009, Acland and Levy 2015), smoking cessation (e.g., Volpp et al. 2009, Giné et al. 2010), and studying (e.g. Barrera-Osorio et al. 2011). In this paper, we shift our focus from incentives to a more compelling form of intervention—forced intervention. We study whether and how forced intervention can promote persistent behavioral change in the adoption of algorithms after the intervention and reduce algorithm aversion among workers during the cold-start stage of algorithm development.

We collaborate with a leading online education platform to conduct our study. The platform provides English classes taught by native English-speaking teachers in the US to Chinese children aged 4 to 15. The primary customer acquisition channel for this company is telephone sales, where sales workers call potential customers (referred to as "leads"), help them book a trial class, and follow up to complete the purchase after the trial. A trial class selection algorithm, developed in late 2018, helps sales workers assign the most appropriate trial class teacher for each lead. The algorithm is designed to maximize post-trial conversion rates while considering teacher availability and student needs. Despite its simplicity and convenience, this algorithm was underutilized: our data show that prior to the experiment (from March 2019 to May 2019), the average algorithm usage ratio, defined as the proportion of trial classes booked using the algorithm, was only 20.08%. Additionally, sales workers tended to use the algorithm for low-quality leads (i.e., those less likely to convert), and the conversion rate for bookings made using the algorithm was significantly lower (9.9% in relative magnitude) than those made manually by the workers during the same period.

To answer our research question, we utilize a randomized field experiment conducted by the company from May 30, 2019 to June 18, 2019 regarding the trial class selection algorithm. During the experiment, a subset of sales workers were randomly assigned into three groups: the Algorithm group, the Manual group, and the Control group. Workers in the Algorithm group were forced to use the algorithm to select teachers for all trial classes, while those in the Manual group were not allowed to use the algorithm and could only manually select trial class teachers.[4] Workers in the Control group did not experience any policy change and were free to select teachers using the algorithm or manually. Notice that if workers chose to use the algorithm for teacher selection, they must accept the algorithm's choice and could not override it by design of the system. After the experiment period, all interventions were removed, and all workers could freely choose between and manual teacher selection. We collect data on sales worker performance for three months before the experiment, during the experiment, and one month after the experiment.

[4] It's important to note that the original intent of this experiment was to evaluate the algorithm's performance, not necessarily to enhance its adoption. Hence, the company included the Manual group.

We provide the first field evidence that forced intervention significantly improves workers' likelihood of using the algorithm after the intervention is removed. Specifically, forcing workers to use the algorithm for three weeks during the experiment significantly increases their post-intervention algorithm usage by 15.8 percentage points compared to the control group. It is important to note that the experiment does not generate significant short-run performance gains because the algorithm performs comparably to human selection; instead, our core focus lies in understanding and improving adoption and belief updating in these "early-stage algorithm" settings. We also show that the post-intervention effect is not driven by a small number of workers but exists across workers with different pre-intervention algorithm usage levels, and it does not decay over time (at least within our observation period). Moreover, we find that forced intervention has no significant impact on the conversion rate of bookings after the experiment. This indicates that our intervention does not adversely affect the firm's operational performance; rather, it may yield long-term benefits by facilitating algorithm adoption and enhancing its performance through increased data collection.

We then investigate the underlying mechanisms through which forced intervention may catalyze sustainable changes in workers' algorithm usage after the intervention. We first build a theoretical model that incorporates two potential underlying mechanisms highlighted in the literature—learning and habit formation—and derive empirical strategies to identify learning considering the possible presence of habit formation. Our empirical evidence suggests that learning exists in the Algorithm group but not in the Control group. In particular, in the Algorithm group, workers who experience a *more* positive change in the conversion rate during the experiment tend to have *higher* algorithm usage after the experiment, but this relationship does not exist in the Control group. Additionally, after the experiment, workers in the Algorithm group are more likely to use the algorithm for high-quality leads compared to before the experiment. Notably, the conversion rate during the experiment in the Algorithm group is close to that of the Control group, contrasting with the earlier evidence that algorithmic bookings had a lower conversion rate than manual bookings. This disparity suggests that the previously observed lower conversion rate of algorithmic bookings was due to selection bias: workers selectively use the algorithm for low-quality leads. It also reveals that workers' initial beliefs about the algorithm were negatively biased. All these results indicate that forced use of the algorithm makes workers learn the unbiased algorithm performance and update their beliefs about it positively, which drives their increased algorithm usage after the intervention.

Our work makes several key contributions to the literature: First, we present field evidence of algorithm aversion during the cold-start stage of an algorithm. In particular, different from existing literature which mostly show users' uniform aversion towards using algorithm, we reveal a distinct form of "selective" algorithm aversion where workers selectively use the algorithm for low-quality cases but avoid it for high-quality cases. Second, we identify a new mechanism driving algorithm aversion: Workers' initial negatively biased beliefs about the algorithm performance drives selective algorithm use, which makes

the algorithm performance appear inferior and, in turn, enhances workers' negative beliefs. This creates a self-reinforcing vicious cycle that discourages algorithm use more broadly. Third, we provide the first field evidence that forced intervention can significantly reduce algorithm aversion and facilitate persistent algorithm use after the intervention among workers. Fourth, we propose an empirical strategy to test the learning mechanism with the possible existence of habit formation, and we further provide evidence that forcing workers to use the algorithm leads to a positive change in their beliefs about the algorithm.

Our findings provide several managerial insights for practitioners. Notably, firms should recognize the possible resistance of workers to using algorithms during the cold-start stage, which is often caused by negatively biased beliefs about the algorithm performance. This resistance can be selective, with workers using the algorithm more on low-quality cases. To promote algorithm adoption, firms can implement extrinsic interventions or provide educational programs to help workers understand the benefits of algorithms and form unbiased beliefs about their capabilities. Even if the algorithm does not initially outperform humans during the cold-start stage, promoting its adoption can still be valuable. It can help break the cycle of biased beliefs and selective algorithm use and expose the algorithm to more diverse and representative use cases. This will also improve the quality of collected data and facilitate more effective algorithm training, which benefits the algorithm development and eventually the operational performance of a company in the long run.

The paper proceeds as follows. In Section 2, we review the relevant literature and discuss our contributions. Section 3 introduces our research context and demonstrates the existence of algorithm aversion in our setting prior to the experiment. Section 4 describes the field experiment design and our data. In Section 5, we present the main results of our experiment, particularly the post-intervention effects on algorithm usage. We then investigate the underlying mechanisms for the post-intervention effects in Section 6. Section 7 concludes the paper.

2. Literature Review

Our paper is closely connected to the growing literature on human–algorithm connection and behavioral interventions. In Sections 2.1 to 2.3, we review the literature on algorithm adoption, adherence to algorithmic recommendations, and operational impacts of algorithm applications, respectively. We then review the literature on behavioral interventions in Section 2.4.

2.1. Algorithm Adoption

Our work directly relates to the literature on human preferences for adopting algorithms versus relying on human judgment and the underlying behavioral drivers of these preferences. A large body of research in this stream focuses on algorithm aversion—the phenomenon that people are averse to using algorithms even if they outperform humans (Dietvorst et al. 2015). Numerous lab studies have demonstrated algorithm aversion across various tasks, including forecasting and prediction (Dietvorst et al. 2015, Dietvorst and Bharti 2020, Turel and Kalhan 2023), diagnosis and evaluation (Longoni et al. 2019,

Reich et al. 2023), decision-making (Dargnies et al. 2024), and generative tasks (Castelo et al. 2019). Some field studies further corroborate this phenomenon in practice (Reis et al. 2020, Knight et al. 2022, Liu et al. 2023). Prior research has identified various factors that may contribute to algorithm aversion, such as perception of inferior algorithm performance (Dietvorst et al. 2015), belief about algorithms' limited capabilities (Longoni et al. 2019, Newman et al. 2020), uncertainty of the use cases (Dai and Singh 2021, Lin et al. 2023), liability concerns and threat to professional status (Reis et al. 2020), psychological distance (Kirshner 2024). The literature suggests multiple strategies to mitigate algorithm aversion and promote algorithm adoption, including allowing for human intervention (Dietvorst et al. 2018, Longoni et al. 2019), reshaping beliefs about algorithm capabilities (Reich et al. 2023, Turel and Kalhan 2023), reframing the task nature (Castelo et al. 2019), and encouraging habitual use (Lin et al. 2023). In contrast, some literature documents algorithm appreciation, where people exhibit a preference for algorithmic decisions or advice over equivalent human input (Gunaratne et al. 2018, Logg et al. 2019, You et al. 2022). Our research is particularly related to this stream of literature, as our empirical setting aligns with the contexts examined in these studies. Specifically, most research in this area assumes (often implicitly) that adopting an algorithm means full acceptance of its output, with no consideration of overriding the algorithmic recommendations after adoption. Our empirical context precisely matches this assumption: Workers must either use the algorithm and accept its teacher selections, or make manual selections by themselves. They cannot choose to use the algorithm but then selectively follow its recommendations.

We contribute to this literature in four aspects. First, we provide field evidence of algorithm aversion in a previously unexplored context—in the workplace during the cold-start stage of an algorithm. Second, we identify a new mechanism for algorithm aversion: the aversion stems from negative and biased beliefs about the algorithm performance, which results in selective use of the algorithm on low-quality leads. This creates a self-reinforcing cycle where the poor results from applying the algorithm to inferior cases further strengthen workers' negative beliefs about the algorithm performance. Third, we introduce forced intervention as an effective approach to mitigate algorithm aversion. We demonstrate that forced use of an algorithm for a period can correct users' biased beliefs about the algorithm and reduce their algorithm aversion. Fourth, previous literature primarily demonstrates the immediate or very short-term performance of the mitigation strategies to improve algorithm adoption, but it remains unclear whether they can promote persistent algorithm usage. Our work addresses this gap by showing the efficacy of forced intervention in promoting sustained algorithm usage among workers.

2.2. Adherence to Algorithmic Recommendations

In practice, algorithms are often implemented as part of a human-in-the-loop system, where the algorithm generates recommendations but humans ultimately decide whether and to what extent they follow them. In this setting, a substantial body of research examines the behaviors of non-adherence to or deviations from the algorithmic recommendations. Such non-adherence behaviors have been observed

in many operational decision-making contexts, including inventory management (Van Donselaar et al. 2010), assortment (Kesavan and Kushwaha 2020, Kawaguchi 2021), pricing (Caro and de Tejada Cuenca 2023, Liang et al. 2023), warehouse operations and logistics (Liu et al. 2021, Sun et al. 2022), and healthcare (Ibanez et al. 2018, Lin et al. 2023). These studies present mixed results on whether non-adherence to algorithmic recommendations improve or harm the operational performance. Non-adherence to algorithmic recommendations may result from various factors. The behavioral drivers of algorithm aversion that hinder algorithm adoption—such as lack of trust in algorithms or self-overconfidence—may also reduce adherence to algorithmic recommendations (Kawaguchi 2021, Karlinsky-Shichor and Netzer 2023). Other behavioral drivers of non-adherence include behavioral biases and cognitive workload (Caro and de Tejada Cuenca 2023, Liang et al. 2023), perception of algorithm complexity (Lehmann et al. 2022), mislearning of the algorithm performance (de Véricourt and Gurkan 2023), etc. Besides, non-adherence to algorithmic recommendations may also arise from technical factors such as system inadequacy and incentive misalignment (Van Donselaar et al. 2010), inaccurate data input (Kwon et al. 2023), and information gaps in the algorithm design (Liu et al. 2021). Several empirical studies have demonstrated the efficacy of some specific methods to improve human adherence to algorithmic recommendations, including providing interpretable performance feedback (Caro and de Tejada Cuenca 2023), incorporating users' own judgments into the algorithm (Kawaguchi 2021), and proactively adjusting algorithm output based on predictions of deviations (Sun et al. 2022). Some theoretical literature also propose to leverage human judgment (Ibrahim et al. 2021, Karlinsky-Shichor and Netzer 2023, Balakrishnan et al. 2025) or take into account human decision bias or non-adherence behaviors (Liu et al. 2021, Chen et al. 2023, Grand-Clément and Pauphilet 2024) in the algorithm design to improve the actual operational performance of smart algorithms across various contexts.

The adherence literature in this section and the adoption literature in Section 2.1 share some similarities but also have many differences. Both streams of literature examine human attitudes and preferences toward algorithmic versus human judgment. The literature reveals some common behavioral factors that can hinder both adoption and adherence, such as distrust in algorithms, overconfidence in human judgment, and risk perception. Both streams have also explored interventions that reshape human beliefs about algorithm performance (such as providing performance feedback) to improve adoption or adherence. However, the two literature streams also differ in several aspects: First, the adoption literature studies the binary preferences between algorithms and humans and generally assumes full adherence to algorithm output after adoption. Conversely, the adherence literature examines post-adoption behaviors where humans can selectively override algorithmic recommendations. Second, the adoption literature are primarily lab studies and focus on identifying behavioral drivers of adoption decisions. In contrast, the adherence literature are largely field studies and, in addition to the behavioral drivers, also identifies technical factors that contribute to non-adherence behavior (e.g., data inaccuracies, misaligned incentives in algorithm design). Third, the adoption literature typically explores behavioral interventions to encourage algorithm adoption, while the adherence literature often investigates technical interventions aimed

at improving the algorithm design and its performance to facilitate adherence to its advice. Notice that these technical interventions are suitable during the mature stage of an algorithm when abundant data is available for its training and continuous improvement, but they may not work well during the cold-start stage when we lack sufficient data for such improvement. In this sense, we contribute to the literature by proposing a new operational lever—forced intervention—as an additional strategy to help positively reshape human beliefs about the algorithm performance and facilitate algorithm adoption, particularly during the cold-start stage, and by providing empirical evidence for its effectiveness in practice through a field experiment.

2.3. Operational Impacts of Algorithm Applications

Our research also contributes to the literature on the operational impacts of deploying smart algorithms in practice. While numerous studies empirically demonstrate the value of adopting novel information technologies across various business sectors (Li et al. 2020, Stamatopoulos et al. 2021, Buell et al. 2022, Han et al. 2024), the operational performance of implementing smart algorithms or AI is more mixed. On the one hand, many studies show that algorithm implementation can improve operational outcomes, including worker productivity (Bai et al. 2022, Knight et al. 2022, Wang et al. 2023a), service efficiency and quality (Lu et al. 2023, Adjerid et al. 2023), production yield (Senoner et al. 2022), and financial performance (Karlinsky-Shichor and Netzer 2023). On the other hand, algorithms may also have various limitations, such as lack of field information (Sun et al. 2022) and input inaccuracy (Kwon et al. 2023), which may lead to inferior algorithm decisions and negatively impact operational performance. Using algorithms may also introduce additional interaction time and information overload, which in turn can reduce operational efficiency (Knight et al. 2022). Furthermore, algorithms can reshape human behavior or decisions when humans collaborate with algorithms on decision-making tasks (Boyacı et al. 2023, de Véricourt and Gurkan 2023) or interact with algorithms to communicate information or decisions (Luo et al. 2019, Tong et al. 2021, Cui et al. 2022, Xu et al. 2023a). Considering these complications, companies should carefully decide the extent and manner of using algorithms in their operations based on their business context and strategic goals (Wang et al. 2023b). Our research contributes to this literature by providing empirical evidence of algorithm performance in a previously unexplored setting and, more importantly, revealing a new mechanism explaining the inferior outcomes of implementing an algorithm in practice: Users with a negative bias toward the algorithm performance will selectively use it on low-quality cases, which skews the operational outcomes.

2.4. Behavioral Intervention

Research in behavioral economics has extensively studied the impacts of extrinsic interventions (mainly financial incentives) on human behaviors in various activities such as exercising, studying, smoking cessation, and alcohol consumption (e.g., Charness and Gneezy 2009, Volpp et al. 2009, Giné et al. 2010, Allcott and Rogers 2014, Milkman et al. 2014, Acland and Levy 2015, Royer et al. 2015, Schilbach 2019).

In operations, a large body of literature examines how various behavioral interventions impact workers' behavior and operational performance in diverse business contexts. Examples of these interventions include financial incentives (e.g., Jiang et al. 2021, Xu et al. 2023b), training programs (e.g., Fisher et al. 2021, Buell et al. 2022), performance feedback (e.g., Song et al. 2018), monitoring (e.g., Pierce et al. 2015, Staats et al. 2017), and staffing and scheduling policies (e.g., Tan and Netessine 2019, Ibanez and Toffel 2020, Xu et al. 2022).

Our paper contributes to this stream of literature in three aspects. First, we investigate the impact of behavioral interventions on a substantive and crucial area in a company's operations nowadays—algorithm adoption. Second, while previous literature primarily studies motivational or incentive-based interventions, we investigate the impact of forced intervention which enforces people to undertake certain behaviors. More importantly, since the intervention forces workers to use the algorithm on all the leads, it eliminates selection bias and enables workers to form unbiased beliefs about the algorithm capabilities. This leads to our third contribution: While existing literature mostly focuses on the habit formation mechanism of behavioral interventions, we propose a method to empirically test the learning mechanism with the possible existence of habit formation. We provide evidence that forced use of an algorithm can make workers update their beliefs about the algorithm positively and facilitate their algorithm adoption as a result.

3. Field Setting and Algorithm Aversion

In this section, we introduce the background of our study prior to the experiment. We first describe our field setting in Section 3.1, and then we present evidence for the existence of algorithm aversion before the experiment in Section 3.2.

3.1. Field Setting

We collaborate with a major online education platform based in China. The platform connects Chinese students aged 4 to 15 with native English-speaking teachers in North America through online classrooms. To take classes, students must first purchase a class package from the company, with prices ranging from 800 to 1800 USD. After purchasing a package, students can book classes with teachers on the platform. The company provides a standardized learning path and prepares teaching materials, so that teachers only need to deliver the classes using these prepared materials.

Since the lump sum payment of a class package is not trivial,[5] the company relies on sales workers to sell the class packages. To facilitate this, each new customer is offered a free trial class before making a purchase decision. A typical sales process proceeds as follows: The company attracts potential customers through marketing activities (e.g., online and offline advertising, referrals, etc.). Through these marketing activities, potential customers can register by providing their contact information. A registered potential

[5] The average educational spending on a child is about 8,139 RMB (1,124 USD) per year in China, according to the 2021 China Household Education Finance Survey Report (`https://ciefr.pku.edu.cn/docs//2023-11/c0966ce10e3d4f58a25001735cc2a18c.pdf`).

customer is called a "lead." Once a lead is registered, the system sends her information to a sales worker. The worker then contacts the lead, offering to book a free trial class for her. If the lead agrees, the worker will book a free trial class for the lead with a specific teacher at a specific time. The lead takes the free trial class with the pre-specified teacher through the online class portal. After the trial class, the same sales worker contacts the lead again to inquire if she is willing to purchase a class package. If the lead agrees, the worker will help her make the purchase.[6] Regardless of the lead's purchase decision, the sales process ends at this stage.

The lead do not have to take classes with the same trial class teacher after purchasing a class package, but anecdotal evidence suggests that the trial class teacher significantly impacts new customers' perceptions about the general value of classes. Since the trial class is the major way for leads to learn the value of classes, it is crucial for workers to find a good teacher for the trial class to convert the lead. Moreover, customers' preferences can be highly heterogeneous, and thus workers need to find not only high-quality teachers but also the best match based on each customer's specific needs, which further increases the complexity of the task. A sales worker's monthly compensation is a low base pay plus the bonus determined by the number of class packages sold by her in the month.[7] Thus, workers have strong incentives to select the best matched teacher for each lead.

Initially, workers need to select a teacher manually by themselves. When booking a free trial class, a sales worker first inputs filtering conditions—including class level, date, time slot, and device—into the system.[8] The system then displays a list of teachers who meet the requirements, along with each teacher's profile, including their photo, tenure at the platform, performance record (e.g., class completion rates and reputation scores), qualifications, and details of their three most recent classes, etc.[9] The sales worker then needs to analyze this information to select a high-quality teacher who fits the customer's preferences. As the company's teacher pool has grown to over 100,000, it has become increasingly challenging for workers to select the best teacher for a lead from an enormous number of teachers.

To address the challenge of manual teacher selection, the company developed a machine learning-based matching algorithm and implemented it on September 4, 2018. The algorithm was trained using the historical data of trial classes, including leads' characteristics, teachers' characteristics, and conversion results (i.e., whether a lead purchases a class package after taking the trial class with the teacher). When a new lead is introduced, the algorithm predicts the conversion probability of the lead for each available teacher based on the lead's and teachers' features, and then selects the teacher with the highest predicted

[6] Notice that customers cannot book a trial class or purchase a class package on their own but have to rely on sales workers to do these.

[7] Generally speaking, the bonus a sales worker gets in a month equals the number of class packages sold by her times a certain coefficient. There are multiple levels of coefficients. When the number of class packages sold by the sales worker is larger, the coefficient is also at a higher level.

[8] The platform offers classes with six levels of difficulty, with each teacher qualified to teach classes of certain levels.

[9] By default, the teachers are listed in descending order of their class completion rates, but workers can also choose to sort them by other criteria such as tenure at the platform.

conversion probability for the lead. Despite the implementation of the algorihm, the company allowed workers to freely choose between using the algorithm or booking manually for each lead (See Figure (6a) in Appendix A for the user interface of workers).[10] By system design, when workers opted to use the algorithm for teacher selection, they must accept the algorithm's choices without the option to override them. In other words, algorithm use entails full compliance with its decisions in our context.

3.2. Algorithm Aversion

Since the algorithm became available, the algorithm usage ratio, defined as the number of trial class bookings where the teachers are selected by the algorithm divided by the total number of trial class bookings made by a sales worker, had been low. From March to May 2019, across all workers in our study, the average algorithm usage ratio was 20.08%, and the median was 13.75%. More surprisingly, algorithmic bookings performed worse than manual bookings. The conversion rate of algorithmic bookings was 9.9% (relative magnitude) lower than that of manual bookings (p-value $= 0.018$).[11] Because workers can freely choose for whom to use the algorithm, it is hard to determine whether the lower conversion rate of algorithmic bookings is due to the algorithm's inferior performance in selecting teachers compared to workers or a selection bias on leads—that is, workers tended to use the algorithm for the leads who are inherently less likely to purchase.

In fact, we do find that workers tended to use the algorithm for low-quality leads. The company defines a binary label *QualityLead* which equals 1 if the lead is of high quality (i.e, more likely to convert) and 0 otherwise.[12] The share of high-quality leads in algorithmic bookings was 3.91 percentage points (absolute difference) lower than in manual bookings (p-value < 0.001), indicating that workers selectively use the algorithm for low-quality leads. This implies that workers may have negative beliefs about the algorithm: they believe that the algorithm performs worse than themselves in selecting trial class teachers. Therefore, they only use the algorithm when they think the lead is of low quality and is not worth their effort of selecting the trial class teacher manually.

Later in Section 6.1, we will make use of experimental data to show that when there is no selective algorithm use, the conversion rate of algorithmic bookings is close to that of manual bookings, indicating that the algorithm does not perform worse than the workers. This result confirms that the lower conversion rate of algorithmic bookings compared to manual bookings before the experiment was due to "selective algorithm aversion": workers selectively use the algorithm for low-quality leads but avoid using it for high-quality leads.

[10] The company did not force workers to use the algorithm for all leads because there was resistance from the salesforce department. As we will show later, the possible underlying reason was that the salesforce department did not trust the algorithm.

[11] Due to the NDA (non-disclosure agreement), we cannot reveal the actual conversion rates. Thus, we only report the relative magnitude of the conversion rate of algorithmic bookings compared to manual bookings and t-test result.

[12] The company has several sources of leads. Leads coming from referral programs or search advertising are usually more likely to convert. Therefore, *QualityLead* is labeled as 1 for leads from these two sources and 0 otherwise.

4. Experiment Design and Data

In this section, we introduce our experiment design and data. Section 4.1 outlines our experiment setting and dataset, and Section 4.2 provides a balance check of the experimental groups.

4.1. Experiment Setting and Data

To evaluate the performance of the algorithm, the company ran a field experiment from May 30, 2019 to June 18, 2019. The workers involved in the experiment were randomly allocated to three treatment groups: the Algorithm group, the Manual group, and the Control group. During the experiment, workers in the Algorithm group were forced to use the algorithm to select the trial class teachers for each lead, and those in the Manual group lost access to the algorithm and had to select the trial class teachers manually. Workers in the Control group could still freely choose to use the algorithm or to select teachers manually, just as before the experiment. Figure 6 in Appendix A shows the user interface for workers in the three conditions.

On May 29 (towards the end of the day), workers assigned to the Algorithm group were notified that "from tomorrow, you must use the algorithm to book the trial class teacher for each lead," and those assigned to the Manual group were notified that "from tomorrow, the algorithm will be disabled and you will need to manually book the trial class teacher for each lead." Neither group was told whether the change was temporary or when it would end. On the morning of June 19, both groups were informed that the change had ended and they could again freely decide whether to use the algorithm or not for each lead. Workers in the Control group received no notifications.

The experiment involved 171 workers. This number was determined to balance statistical power and the cost of running the experiment. The participating workers were dispersed across the company's four sales regions and multiple offices within each region, with no more than 10% of the workers in any single office participating (see Table 8 in Appendix A for the distribution of workers across offices).[13] This design minimizes the possibility of workers in different treatment groups interacting and discuss the treatments (i.e., the water cooler effect). We also conduct Fisher's exact tests and show that the three treatment groups do not have significantly different distributions of workers across sales regions (p-value $= 0.882$) and across sales offices (p-value $= 0.697$).

We collect booking-level data from bookings made by the participating workers over three periods: three months before the experiment (March 1-May 29, 2019), during the experiment (May 30-June 18, 2019), and one month after the experiment (June 19-July 17, 2019).[14] A booking refers to a trial class booked for a lead, and we take leads and bookings as having one-to-one correspondence throughout the paper. For each booking, we collect data on two key dependent variables for our analysis: The focal

[13] We cannot report the exact total number of workers in each office due to data confidentiality.

[14] Due to a technology issue, the experiment started in the middle of the day on May 30, 2019, and the Algorithm group and the Manual group did not start at the same time. Because some of our analysis is at a daily level, we drop the data of May 30, 2019 from our analysis. In other words, the pre-experiment period ends on May 29 and the experiment period begins on May 31 in our analysis.

dependent variable is *UseAlgorithm*, which equals 1 if a trial class booking is made using the algorithm and 0 otherwise. Another dependent variable is *Purchase*, which equals 1 if the booking converts to a purchase within 17 days and 0 otherwise.[15] We also collect data on the experimental group assignment of a worker and several control variables. These control variables include booking date (*BookDate*), seniority of a worker (*Seniority*), and several lead characteristics, including student age (*StudentAge*), whether the lead is high-quality or not (*QualityLead*), and whether the booking for the lead is made through a phone call or WeChat (*Call*). Table 7 in Appendix A summarizes the definitions of the variables that we use throughout the paper.

It is important to note that the company included the Manual group in the experiment because the original purpose of the experiment was to evaluate the algorithm performance rather than to promote its adoption. However, for the purpose of our study, we are primarily interested in the impact of forced algorithm use on workers' long-term algorithm adoption behavior. Therefore, we will focus our analysis on the comparison between the Algorithm group and the Control group for the remainder of our paper. Some additional analysis results involving all three experimental groups are provided in Appendix A.

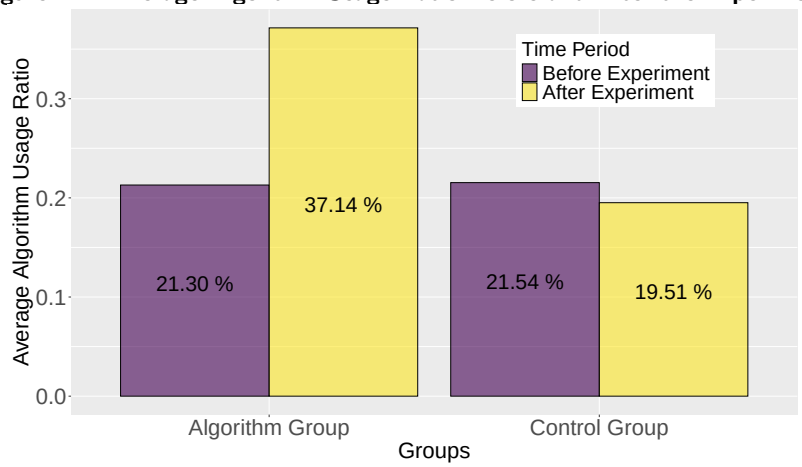
4.2. Balance Check

Table 9 in Appendix A presents a balance check between the Algorithm and Control groups. Panel A shows summary statistics of the booking-level variables, including two dependent variables (*UseAlgorithm* and *Purchase*) and lead characteristics (*StudentAge*, *QualityLead*, and *Call*), between the two groups before the experiment (from March 1, 2019 to May 29, 2019). Panels B and C compare the worker seniority (*Seniority*), number of bookings made by a worker per day (*NumBooking*), and number of converted bookings (i.e., bookings converted into purchases within 17 days) made by a worker per day (*NumPurchase*) between the two groups before the experiment. We report the mean and standard deviation of each variable in each group. We also conduct t-tests to compare the variable means between the two groups. The results show that there are no statistically significant differences between the two groups for any variables. This suggest that the Algorithm and Control groups are generally balanced regarding workers' behavior and the characteristics of the leads and workers before the experiment. We also provide summary statistics during (Table 10) and after the experiment (Table 11) in Appendix A. The lead characteristics remain balanced between the two groups during and after the experiment.[16] These results collectively indicate that any observed differences during and after the experiment between the two groups are not due to pre-existing differences in the workers or leads and should be attributed to the experimental group assignment, i.e., whether a worker is forced to use the algorithm or not. In

[15] According to historical data, the time between booking a trial class and the corresponding purchase has a large variation, with 17 days as the 90th percentile. Thus we use 17 days as the cut-off when defining *Purchase* so that the comparison between the conversion rates of early and late bookings is fair.

[16] The company has a system to allocate leads to each sales worker, which balances fairness and efficiency. The workers cannot decide which leads they receive. We cannot make leads allocation purely random during the experiment, but our balance checks make sure that the leads allocated to different experimental groups do not have significant differences in the key characteristics.

Figure 1 **Average Algorithm Usage Ratio Before and After the Experiment**



Appendix B.1, we further show that our main empirical results are not influenced by workers' exiting behavior.

5. Post-intervention Treatment Effects

In this section, we investigate how forced intervention impacts workers' algorithm adoption and operational performance after the experiment. We first evaluate the effects of forced intervention on workers' post-intervention algorithm usage in Section 5.1 and examine the robustness of these effects across worker groups and over time in Section 5.2. We then analyze how forced intervention impacts the operational performance in Section 5.3.

5.1. Post-intervention Algorithm Usage

Figure 1 compares the average algorithm usage ratios, i.e., the means of *UseAlgorithm*, before and after the experiment between the Algorithm and Control groups. After the experiment, the average algorithm usage ratio increases by 15.8 percentage points (p-value < 0.001) for the Algorithm group, which is nearly doubled, while it does not change significantly for the Control group (p-value $= 0.550$). The post-intervention average algorithm usage ratio of the Algorithm group is significantly higher than that of the Control group (p-value $= 0.002$).[17]

We next perform a Difference-in-Differences (DID) analysis to estimate the impact of forced intervention on the algorithm usage ratio after the experiment. Specifically, for all bookings made by the workers in the Algorithm and Control groups before and after the experiment, we conduct an ordinary least square (OLS) regression analysis using the following specifications:

$$UseAlgorithm_{ijt} = \beta_0 + \beta_1 AlgorithmGroup_i \times AfterExp_t + \beta_2 AlgorithmGroup_i + \beta_3 AfterExp_t + \epsilon_{ijt}, \quad (1)$$

$$UseAlgorithm_{ijt} = \beta_0 + \beta_1 AlgorithmGroup_i \times AfterExp_t + \tau_t + \mu_i + X_j + \epsilon_{ijt}. \quad (2)$$

[17] All comparisons here are t-tests, with the standard errors clustered at the worker level.

Table 1 Post-Intervention Effect on Algorithm Usage

	Dependent variable: *UseAlgorithm*			
	(1)	(2)	(3)	(4)
AlgorithmGroup × *AfterExp*	0.1787***	0.1803***	0.1959***	0.1965***
	(0.0543)	(0.0540)	(0.0534)	(0.0532)
AlgorithmGroup	-0.0024	-0.0008		
	(0.0372)	(0.0372)		
AfterExp	-0.0202			
	(0.0336)			
Date Fixed Effects	No	Yes	Yes	Yes
Worker Fixed Effects	No	No	Yes	Yes
Lead Characteristics	No	No	No	Yes
Relative Effect Size	83.0%	83.7%	90.9%	91.2%
Observations	71,644	71,644	71,644	71,644

Note: This table reports the treatment effects of forced intervention on post-intervention algorithm usage. Column (1) shows the results without controlling for any fixed effects and lead characteristics. Columns (2)-(4) sequentially add date fixed effects, worker fixed effects, and lead characteristics. The pre-intervention algorithm usage ratio of the Control group (21.54%) is used as the baseline to compute relative effect sizes. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

In Equations (1) and (2), $UseAlgorithm_{ijt}$ is a binary variable that equals 1 if lead j's trial class booking made by worker i on day t was made using the algorithm and 0 otherwise. $AlgorithmGroup_i$ is a binary variable that equals 1 if worker i is in the Algorithm group and 0 otherwise. $AfterExp_t$ is a binary variable that equals 1 if day t is after the experiment and 0 otherwise. τ_t and μ_i represent the date fixed effects and worker fixed effects, respectively. X_j represents characteristics of lead j, including $StudentAge_j$, $QualityLead_j$, and $Call_j$ as defined in Section 4.1. The coefficient of the interaction term β_1 measures the post-intervention effect on algorithm usage for the Algorithm group compared to the Control group.

Table 1 presents the estimation results. We find that, for the Algorithm group, the intervention of forcing workers to use the algorithm significantly increases their post-intervention algorithm usage compared to the Control group by 17.9 percentage points. This corresponds to a relative increase of 83% in the post-intervention algorithm usage, with the pre-intervention average algorithm usage ratio in the Control group (21.54%) as the baseline. Adding date fixed effects, worker fixed effects, or lead characteristics does not change the estimates of the post-intervention effect in the Algorithm group significantly, as shown in Columns (2)-(4). The treatment effects on the individual-level algorithm usage ratio are also similar, as presented in Appendix B.2. In Appendix B.3, we also examine the heterogeneous treatment effects of forced intervention across workers, and we find no significant heterogeneity in its impact on post-intervention algorithm usage across workers with different seniority or past performance measures.

5.2. Robustness of Post-intervention Effects on Algorithm Usage

In this section, we examine the robustness of the post-intervention effects of forced intervention on algorithm usage across different worker groups and over time. We first explore whether the change in algorithm usage in the Algorithm group is driven by a small number of workers who had rarely used the algorithm before the experiment or if it is relatively prevalent among the workers. Note that only four out of the 171 workers had never used algorithm before the experiment, and only two of them are in the Algorithm group. Therefore, the significant increase in average algorithm usage in the Algorithm

Figure 2 Algorithm Usage Before and After the Experiment by Worker Groups

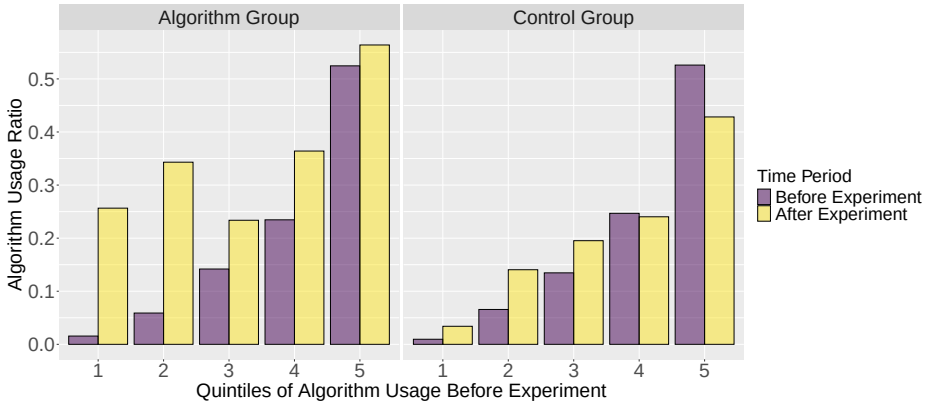


Table 2 Post-intervention Change in Algorithm Usage by Worker Groups

	Quintiles of Algorithm Usage Before Experiment				
Groups	1	2	3	4	5
Algorithm group	0.241**	0.284***	0.092**	0.130	0.039
	1547.5%	482.1%	64.7%	55.2%	7.5%
Control Group	0.024**	0.075	0.061	-0.007	-0.098
	255.7%	114.0%	45.0%	-2.7%	-18.6%

Note: Quintile n refers to the workers whose pre-experiment algorithm usage ratio falls within the n-th quintile among all workers. For each experimental group, the first row reports the absolute change in algorithm usage ratio (post-intervention usage ratio − pre-intervention usage ratio) within each quintile, and the second row reports the relative change in algorithm usage ratio ((post-intervention usage ratio − pre-intervention usage ratio)/pre-intervention usage ratio) within each quintile. For each quintile in each group, we conduct a t-test to compare the algorithm usage ratio at the worker level before and after the experiment. The stars in the first row of each group indicate the significance of the t-tests. *p<0.1; **p<0.05; ***p<0.01.

group cannot be attributed to those who did not use the algorithm before but started to use it after the experiment. To further investigate the distribution of the changes in algorithm usage among workers, we group the workers into five quintiles according to their pre-experiment algorithm usage ratios. Quintile 1 includes the workers whose pre-experiment algorithm usage ratios are among the lowest 20 percent out of all workers. Quintile 2 includes those with pre-experiment algorithm usage ratios between the 20th and 40th percentiles, and so on. Figure 2 and Table 2 show the average algorithm usage ratios before and after the experiment and their comparisons for each quintile in the Algorithm and Control groups. For the Algorithm group, all the five quintiles have higher average post-intervention algorithm usage compared to pre-intervention levels, with the changes in algorithm usage for the first three quintiles being statistically significant. These findings demonstrate that the change in algorithm usage is not driven by a small number of workers but is relatively prevalent across workers with various levels of pre-intervention algorithm usage, though the magnitude of the change varies depending on their pre-intervention usage levels.

We next investigate whether the change in algorithm usage is driven by only a few days after the experiment or remains relatively stable over time after the experiment. Figure 3 shows the average algorithm usage ratio of bookings on each day before, during, and after the experiment. We test the

Figure 3 Daily Average Algorithm Usage Ratio

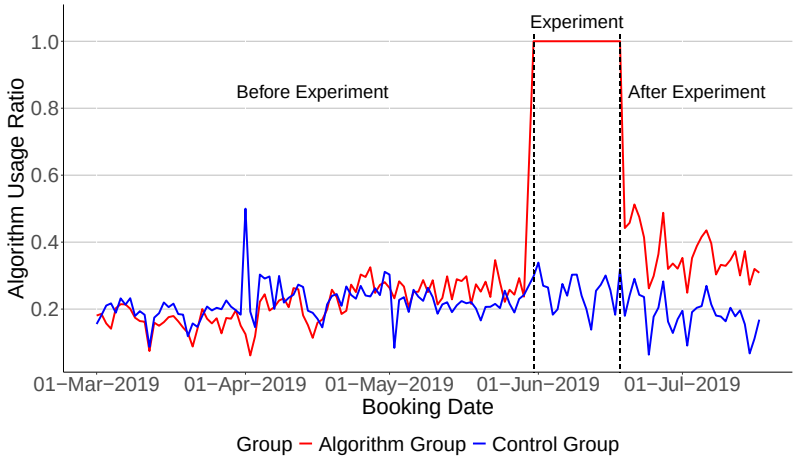


Table 3 Trend of Algorithm Usage Before and After Experiment

	Dependent variable: *UseAlgorithm*
AlgorithmGroup × *AfterExp* × *DayToExperiment*	-0.0034
	(0.0023)
AlgorithmGroup × *AfterExp*	0.1677***
	(0.0568)
AlgorithmGroup × *DayToExperiment*	0.0013*
	(0.0007)
Date Fixed Effects	Yes
Worker Fixed Effects	Yes
Lead Characteristics	Yes
Observations	71,644

Note: This table reports the estimation results on the trend of algorithm usage before and after experiment. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

trend of algorithm usage before and after the experiment in the Algorithm group with Control group as the benchmark using the following OLS regression specification:

$$\begin{aligned} UseAlgorithm_{ijt} =& \gamma_0 + \gamma_1 AlgorithmGroup_i \times AfterExp_t + \gamma_2 AlgorithmGroup_i \times DayToExperiment_t \\ &+ \gamma_3 AlgorithmGroup_i \times AfterExp_t \times DayToExperiment_t + \tau_t + \mu_i + X_j + \epsilon_{ijt}. \end{aligned} \tag{3}$$

In Equation (3), $DayToExperiment_t$ is defined as the difference in days between day t and May 30, 2019 for days before the experiment (taking negative values) and the difference in days between day t and June 18, 2019 for days after the experiment (taking positive values). The other variables are defined the same as those in Equation (2). Table 3 presents the estimation results of Equation (3). The coefficient of *AlgorithmGroup* × *AfterExp* remains significantly positive with a similar magnitude as those in Table 1, while the coefficient of the three-way interaction term *AlgorithmGroup* × *AfterExp* × *DayToExperiment* is not statistically significant. This suggests that the post-intervention improvement in algorithm usage in the Algorithm group is robust after accounting for the pre- and post-intervention trends and remains stable after the experiment.

Table 4 Post-Intervention Effect on Operational Performance

Dependent variable:	*Purchase*		*NumBooking*		*NumPurchase*	
	(1)	(2)	(3)	(4)	(5)	(6)
AlgorithmGroup × *AfterExp*	0.0063	0.0011	−0.0831	−0.0616	0.0263	0.0066
	(0.0053)	(0.0047)	(0.2249)	(0.1740)	(0.0358)	(0.0311)
AlgorithmGroup	0.0015		−0.4128		−0.0069	
	(0.0046)		(0.2737)		(0.0331)	
AfterExp	0.0187***		0.0334		0.1252***	
	(0.0032)		(0.1573)		(0.0231)	
Date Fixed Effects	No	Yes	No	Yes	No	Yes
Worker Fixed Effects	No	Yes	No	Yes	No	Yes
Lead Characteristics	No	Yes	No	No	No	No
Observations	71,644	71,644	11,206	11,206	11,206	11,206

Note: This table reports the estimated treatment effects on conversion rate of bookings (Columns (1)-(2)), number of bookings made by a sales worker per day (Columns (3)-(4)) and number of bookings converted into purchases per worker per day (Columns (5)-(6)). Columns (1), (3) and (5) do not control for date and worker fixed effects or lead characteristics. Column (2) controls for date and worker fixed effects and lead characteristics. Columns (4) and (6) control for date and worker fixed effects. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

5.3. Post-intervention Operational Performance

We finally investigate how forced intervention impacts workers' operational performance after the experiment. We focus on three operational outcomes: conversion rate of all bookings, total number of bookings made by a worker per day, and number of bookings converted into purchases per worker per day. To this end, we examine the three dependent variables: (1) $Purchase_{ijt}$, which equals 1 if lead j's trial class booking made by worker i on day t converts to a purchase within 17 days and 0 otherwise, (2) $NumBooking_{it}$, which represents the total number of bookings made by worker i on day t, and (3) $NumPurchase_{it}$, which represents the number of converted bookings (i.e., bookings converted into purchases within 17 days) made by worker i on day t. For the analysis of conversion rate of all bookings, we perform OLS regression analyses as in Specifications (1) and (2) with $Purchase_{ijt}$ as the dependent variable. For the analysis of number of bookings and converted bookings made by a worker per day, we use the following OLS regression specifications:

$$Outcome_{it} = \lambda_0 + \lambda_1 AlgorithmGroup_i \times AfterExp_t + \lambda_2 AlgorithmGroup_i + \lambda_3 AfterExp_t + \epsilon_{it}, \quad (4)$$

$$Outcome_{it} = \lambda_0 + \lambda_1 AlgorithmGroup_i \times AfterExp_t + \tau_t + \mu_i + \epsilon_{it}, \quad (5)$$

where $Outcome_{it} \in \{NumBooking_{it}, NumPurchase_{it}\}$.

The estimation results are presented in Table 4. We find that forcing workers to use the algorithm does not significantly impact the three operational performance measures after the experiment.[18,19] This

[18] The conversion rate and number of converted bookings per worker per day directionally increase, suggesting a potential improvement in the conversion performance. The total number of bookings per worker per day directionally decreases. Since booking trial classes is only part of the workers' daily responsibilities, this decrease primarily implies a potential change in workers' effort allocation across their duties rather than a decrease in their overall productivity.

[19] We conduct a power analysis showing that the sample size for the individual-level analysis in Table 4 is sufficient to detect small-to-moderate treatment effects, and thus the insignificant interaction terms in these columns likely reflect the treatment effects being absent or too small to be detected rather than a result of insufficient sample size. Please see Appendix B.4 for more details.

suggests that while the workers use the algorithm more frequently after the experiment, it does not lead to substantial changes in their booking conversions and the effort allocated to making bookings. We also plot the daily average values of the three metrics for both groups over time in Figure 7 in Appendix A, which further illustrates the absence of significant changes in the post-intervention operational performance. The lack of improvement in the conversion rate is due to the fact that the algorithmic bookings actually achieve similar (or weakly superior) conversion performance compared to manual bookings, as we will show in Section 6.1. Therefore, switching from manual bookings to algorithmic bookings for more leads does not lead to notable improvements in the conversion rate. From the company's perspective, this is favorable: Increased algorithm usage during its cold-start stage does not adversely affect the operational outcomes of the company, even though the algorithm performance is not yet superior. However, it can facilitate data collection and thus contribute to iterative improvement of the algorithm performance in the future. This will in turn benefit the company's operational performance in the long run.

6. Mechanisms

In this section, we investigate the possible underlying mechanisms for the post-intervention increase in algorithm usage as shown in Section 5. As discussed in Section 1, prior research (e.g., Charness and Gneezy 2009, Acland and Levy 2015) has documented that providing extrinsic incentives (e.g., monetary incentives) for specific behaviors (e.g., gym attendance) can induce persistence of such behaviors after the incentives are removed. The literature identifies two primary mechanisms explaining such post-intervention behavioral changes: *learning* and *habit formation*. Learning means that people learn the benefit of a behavior through experience during the intervention period and thus continue the behavior later. Habit formation means that people have formed the habit of a behavior during the intervention period, which persists after the intervention.

In our setting, workers are forced, rather than incentivized, to use the algorithm for a specific period. However, the underlying principle remains similar: workers' algorithm usage behavior is altered by an extrinsic intervention for a period of time. Thus, following the existing literature, we also consider learning and habit formation as two possible mechanisms through which forced algorithm use may change workers' post-intervention algorithm usage behavior. Notice that previous studies mostly interpret their findings as consistent with habit formation but not with learning. In contrast, our focus is to identify whether the learning mechanism exists considering the possible presence of habit formation, which, as far as we know, has not been empirically tested in existing literature. We incorporate the habit formation mechanism into our theoretical illustration to analyze how we can disentangle learning from habit formation.

This section proceeds as follows: In Section 6.1, we first discuss the possibility of the existence of the learning mechanism by examining the unbiased performance of the algorithm, using the data during the experiment. Then in Section 6.2, we develop a theoretical model that incorporates both learning and habit formation mechanisms and derive the testing rule to identify learning with the possible existence of habit formation . Finally, in Section 6.3, we empirically test the existence of learning based on the derived testing rule using our data.

6.1. Unbiased Performance of the Algorithm

During the experiment, workers in the Algorithm group were forced to use the algorithm for all their bookings. We have also shown in Section 4.2 that there is no significant difference in lead characteristics between the Algorithm and Control groups during the experiment. Thus, comparing conversion rates between these two groups during the experiment can provide an unbiased assessment of the algorithm performance compared to manual bookings. As shown in Table 13 in Appendix A, the Algorithm group has a directionally higher conversion rate than the Control group during the experiment, though the difference is not statistically significant. The insignificant improvement of the algorithm may be explained by the algorithm's information limitations: While the algorithm has an advantage over workers in that it has complete information on available teachers and their characteristics, it has very limited information on lead characteristics. In contrast, workers can know more about the leads' preferences through conversations with them and incorporate these insights when selecting trial class teachers for new leads. Recall that before the experiment, the conversion rate of algorithmic bookings was significantly lower than that of manual bookings (see Section 3.2). This unbiased performance assessment of the algorithm indicates that the lower conversion rate of algorithmic bookings before the experiment was due to selective algorithm adoption, i.e., workers tended to use the algorithm for lower-quality leads.

Given the results above, learning could work as follows: Before the experiment, workers underestimate the algorithm's ability to select the best trial class teacher, so they tend to underutilize the algorithm and use it for low-quality leads. This behavior in turn reinforces their negative beliefs about the algorithm and thus enhances their algorithm aversion. During the experiment, workers in the Algorithm group are forced to use the algorithm on all bookings and get exposed to the unbiased algorithm performance. Even if they find that the algorithm performance is just similar to manual bookings, it is still an improvement compared to their prior belief that the algorithm has a significantly lower conversion rate than manual selections. In fact, the conversion rate of the algorithmic bookings in the Algorithm group is 2.0 percentage points higher during the experiment than before the experiment. As a result, workers in the Algorithm group may update their beliefs about the algorithm from a negatively biased perspective to an unbiased one, which reduces their algorithm aversion. Therefore, workers in the Algorithm group are more likely to use the algorithm after the experiment because they have developed a more positive belief about its capabilities.

However, the increase in algorithm usage after the experiment does not necessarily indicate the presence of learning. As noted earlier, habit formation can also contribute to the observed post-intervention effects: forcing workers to use the algorithm for a period makes them develop the habit of using it. Disentangling the learning mechanism from habit formation is challenging, since they lead to similar post-intervention effects and the change in an individual's belief is difficult to measure. However, unlike previous studies, we benefit from access to more detailed data, such as lead quality and conversion rates, which enables us to infer changes in workers' beliefs. In what follows, we develop a simple illustrative model that

incorporates both learning and habit formation and derive the testing rule to identify the presence of learning while accounting for the possible existence of habit formation. We then empirically test the learning mechanism.

6.2. Theoretical Illustration

Suppose there are three time periods $t=0$ (before the experiment), 1 (during the experiment), 2 (after the experiment). In each period t, each sales worker i chooses $a_{ti} \in [0,1]$, the probability of using the algorithm when booking the trial class for each lead. The Algorithm and Control groups are denoted as A and C, respectively. We assume a worker i's expected utility from each lead in time period t has the following functional form:

$$u(a_{ti}) = R(a_{ti}) - C(a_{ti}) - H(a_{ti} - a_{t-1,i}), \tag{6}$$

where R, C, and H capture the reward, effort cost and habit-related cost, respectively. For clarity and tractability of the theoretical illustration, we next derive our testing rule using some simple functional forms of R, C and H. In Appendix C, we show that our testing rule remains valid for more general functional forms (under certain regularity assumptions) and at the daily time scale for period t.

The first part $R(a_{ti})$ is the expected reward from a lead given the probability of using the algorithm a_{ti}. We assume that a worker receives a monetary reward r_0 if a lead is converted. Suppose the conversion rate is w for manual bookings and $(1+v)w$ for algorithmic bookings, where v measures the relative performance of algorithmic bookings compared to manual bookings. Assume workers are risk-neutral. Then we can express the expected reward from a lead as $R(a_{ti}) = r_0[(1+v)wa_{ti} + w(1-a_{ti})] = r(1+va_{ti})$, where $r \equiv r_0 w$ is the expected reward of a manual booking.

The second part $C(a_{ti})$ is the effort cost associated with serving a lead when the probability of using the algorithm is a_{ti}. We assume that $C(a_{ti}) = \frac{1}{2}[(c+c_a)a_{ti}^2 + c(1-a_{ti})^2]$, where c is the cost parameter for manual bookings and $c_a \in (-c, \infty)$ captures the difference in the cost parameter between algorithmic bookings and manual bookings. For simplicity here, we assume that c and c_a are homogeneous across workers and leads. With the algorithm usage probability being a_{ti}, a worker decides whether to use the algorithm for each lead based on her assessment of the lead's suitability for algorithmic booking. The quadratic functional form of $C(a_{ti})$ captures the increasing marginal effort cost as a_{ti} increases: As workers expand algorithm usage on additional leads, the marginal savings in booking effort (i.e., effort spent on selecting teachers) decreases, since workers tend to automate the most time-consuming manual bookings first. However, the marginal decision effort increases, as it becomes increasingly difficult to determine on borderline cases whether relying on the algorithm or personal judgment is better. Workers will balance algorithmic bookings and manual bookings based on the cost parameters. The most cost-efficient algorithm usage probability is reversely proportional to the cost parameter, i.e., $a_{ti}/(1-a_{ti}) = c/(c+c_a)$. A lower c_a means that algorithmic bookings are more cost-saving and should lead to a higher algorithm usage probability.

The third part $H(a_{ti} - a_{t-1,i})$ represents the disutility of changing the algorithm usage compared to the last time period. Habit formation means that people experience disutility when deviating from their established states (Becker and Murphy 1988, Becker 1998, Charness and Gneezy 2009). We assume that $H(a_{ti} - a_{t-1,i}) = h \cdot (a_{ti} - a_{t-1,i})^2$, where $h > 0$ indicates the presence of habit formation. This functional form means that the disutility increases with the absolute difference between a_{ti} and $a_{t-1,i}$. In other words, a higher $a_{t-1,i}$ leads the worker to prefer a higher a_{ti} as well.

To summarize, worker i's expected utility from each lead in time period t is:

$$u(a_{ti}) = r(1 + va_{ti}) - \frac{1}{2}[(c + c_a)a_{ti}^2 + c(1 - a_{ti})^2] - h(a_{ti} - a_{t-1,i})^2. \tag{7}$$

We first consider the time period before the experiment ($t = 0$). Since workers have been using manual bookings and have also been exposed to the algorithm for a relatively long time, we assume that they already know the exact values of the cost parameters for manual bookings (c) and algorithmic bookings ($c + c_a$), as well as the expected reward of manual bookings (r). Due to the selective algorithm adoption for certain leads, workers do not know exactly the relative performance of the algorithm compared to manual bookings (v), but they have formed a belief $v \sim N(\mu_{0i}, \sigma_{0i}^2)$. In Table 14 in Appendix A, we provide evidence that there is no learning during our observation period before the experiment. Therefore, it is reasonable to assume that workers have formed a relatively stable belief about the performance of algorithm and manual bookings before the experiment. We also assume that the pre-experiment period is long enough so that workers have reached a steady state, i.e., the choice of a_{0i} is relatively stable (which is also consistent with the empirical evidence shown in Figure 3). Then the term $h(a_{ti} - a_{t-1,i})^2$ is 0 here. Therefore, before the experiment, the optimal choice of a_{0i} should maximize $\mathrm{E}[u(a_{0i})] = r(1 + \mu_{0i}a_{0i}) - \frac{1}{2}[(2c + c_a)a_{0i}^2 - 2ca_{0i} + c_a]$. Solving the first-order condition, we get worker i's algorithm usage probability in $t = 0$ of the two groups:

$$a_{0i} = \frac{r\mu_{0i} + c}{2c + c_a}, i \in A, C. \tag{8}$$

Next, we consider the periods during the experiment ($t = 1$) and after the experiment ($t = 2$). We discuss the Algorithm and Control groups separately. For workers in the Algorithm group, they are forced to use the algorithm for every lead during the experiment, i.e., $a_{1i} = 1$ for $i \in A$, and they may update their beliefs about v. Notice that different workers may have different prior beliefs and update their beliefs differently. We assume that worker i's updated belief is $v \sim N(\mu_{1i}^A, (\sigma_{1i}^A)^2)$ upon the end of the experiment, and she determines her post-experiment algorithm usage probability a_{2i} based on this updated belief. Thus, worker i chooses a_{2i} to maximize $\mathrm{E}[u(a_{2i})] = r(1 + \mu_{1i}^A a_{2i}) - \frac{1}{2}[(2c + c_a)a_{2i}^2 - 2ca_{2i} + c] - h(a_{2i} - a_{1i})^2$. Solving its first-order condition, we can derive that

$$a_{2i} = \frac{r\mu_{1i}^A + c + 2h}{2c + c_a + 2h}, i \in A. \tag{9}$$

For workers in the Control group, given that they have reached a steady state before the experiment and nothing changes during or after the experiment, it is reasonable to assume that their belief about v

remains constant across the three periods and their steady-state algorithm usage probability also remains unchanged, i.e., $a_{2i} = a_{1i} = a_{0i}$ for $i \in C$. This is also supported by our data: As shown in Figure 3, the algorithm usage among workers in the Control group does not change significantly before, during or after the experiment. Statistical tests also confirm that the average algorithm usage ratios during and after the experiment are not significantly different from that before the experiment. Specifically, the change in the algorithm usage ratio during the experiment compared to before is 0.039 with p-value $= 0.150$, and the change after the experiment compared to before is -0.020 with p-value $= 0.550$.

Now we consider how to identify learning with the possible presence of habit formation. Learning means that workers update their beliefs about the relative performance of the algorithm v, which will cause a change in the mean belief (i.e., $\mu_{1i} - \mu_{0i} \neq 0$) and a reduction in uncertainty ($\sigma_{1i} < \sigma_{0i}$). Given the assumption of risk-neutrality, σ_{ti} does not affect workers' choices of a_{ti}, and thus we focus on the change in the mean belief μ_{ti} hereafter. If learning does not exist, each worker's belief remains unchanged, i.e., $\mu_{1i} = \mu_{0i}$. The existence of habit formation means $h > 0$.

We calculate the change in the algorithm usage ratio $\Delta a_i = a_{2i} - a_{0i}$ for the Algorithm group:

$$\Delta a_i = \frac{r(\mu_{1i}^A - \mu_{0i})}{2c + c_a + 2h} + 2h \frac{c + c_a - r\mu_{0i}}{(2c + c_a)(2c + c_a + 2h)}, \text{ for } i \in A. \tag{10}$$

Equation (10) indicates that if learning exists, Δa_i increases with $\mu_{1i}^A - \mu_{0i}$, and this relationship does not depend on whether habit formation exists or not. If there is no learning ($\mu_{1i}^A - \mu_{0i} = 0$) but habit formation exists ($h > 0$), the algorithm usage ratio should increase ($\Delta a_i > 0, i \in A$).[20] However, an increase in the algorithm usage ratio in the Algorithm group alone does not imply the existence of habit formation, because it could also be driven by an increase in the mean belief ($\mu_{1i}^A - \mu_{0i} > 0$). To test the learning mechanism, the key is how to measure $\mu_{1i}^A - \mu_{0i}$ empirically. Recall that μ_{1i}^A and μ_{0i} are the expected values of worker i's belief about the relative performance of algorithmic bookings on the conversion rate compared to manual bookings (v) during and before the experiment, respectively. Intuitively, workers can update their beliefs about v based on the change in the conversion rate when they are forced to use the algorithm on all bookings. Thus, we measure $\mu_{1i}^A - \mu_{0i}$ using the change in the conversion rate that worker i experiences from the pre-experiment period $t = 0$ to the experiment period $t = 1$, i.e., we measure $\mu_{1i}^A - \mu_{0i}$ using $\bar{\theta}_{1i} - \bar{\theta}_{0i}$, where $\bar{\theta}_{ti}$ represents the conversion rate of bookings made by worker i during period t.[21]

[20] Given $a_{0i} < 1$, we know that $c + c_a - r\mu_{0i} > 0$. Then for $i \in A$, $\Delta a_i = \frac{2h(c + c_a - r\mu_{0i})}{(2c + c_a)(2c + c_a + 2h)} > 0$.

[21] Notice that for workers in the Algorithm group, $\bar{\theta}_{1i} = \bar{\theta}_{1i}^A$, i.e., the conversion rate of worker i in $t = 1$ is the conversion rate of solely algorithmic bookings made by worker i in $t = 1$, whereas $\bar{\theta}_{0i}$ is a linear combination of the conversion rates of manual and algorithmic bookings made by worker i in $t = 0$. An alternative measure for $\mu_{1i}^A - \mu_{0i}$ is the change in the conversion rate of algorithmic bookings that worker i experienced from $t = 0$ to $t = 1$, i.e., $\bar{\theta}_{1i}^A - \bar{\theta}_{0i}^A$, where $\bar{\theta}_{ti}^A$ is the conversion rate of algorithmic bookings made by worker i during period t. Between the two measures for $\mu_{1i}^A - \mu_{0i}$, we use $\bar{\theta}_{1i} - \bar{\theta}_{0i}$ in our main analysis for two reasons: First, most workers used the algorithm for only a small number of bookings before the experiment, and thus $\bar{\theta}_{0i}^A$ can be a noisy measure for μ_{0i}. Second, for the same reason, workers probably had a clearer perception of their overall conversion rate than the specific conversion rate of algorithmic bookings before the experiment. That said, we also use the alternative measure $\bar{\theta}_{1i}^A - \bar{\theta}_{0i}^A$ to test the learning mechanism and present the results in Table 16 in Appendix A. The results are qualitatively similar to those in Section 6.3, though the coefficient is smaller and less significant, which is possibly driven by the noisy nature of $\bar{\theta}_{0i}^A$ as explained earlier.

Notice that the same sales worker helps a lead book the trial class and make the purchase, so workers know whether each booking they make has converted or not. Although we use 17 days as the cutoff to define a purchase, over 50% of purchases actually occur within 2 days after the trial class booking, and over 75% of purchases occur within 6 days after the trial class booking. This means that the workers should be able to perceive the algorithm performance within a few days after the trial class bookings in most cases and thus have sufficient time during the experiment period to learn about the algorithm performance. Thus, we assume that the duration of the experiment is long enough for workers to understand how the intervention affects their conversion rates.

Based on the analysis above, we derive the following rule to test the learning mechanism, which does not depend on whether or not habit formation exists: For the Algorithm group, if the change in a worker's algorithm usage ratio (from the pre-intervention period to the post-intervention period) *increases* with the change in her experienced conversion rate (from the pre-intervention period to the intervention period), it indicates that learning exists. If these two changes are uncorrelated, it indicates that learning does not exist. Note that if learning exists, we cannot separately identify whether habit formation exists based on the increase in algorithm usage in the Algorithm group. However, if learning does not exist, then the increase in algorithm usage in the Algorithm group indicates the existence of habit formation. In fact, identifying habit formation when learning exists presents inherent challenges, and prior studies on habit formation (e.g., Charness and Gneezy 2009, Acland and Levy 2015) typically assume no learning or acknowledge learning as an alternative mechanism without distinguishing the two mechanisms. The challenges arise since both habit formation and learning lead to similar observable outcomes (i.e., sustained behavior after external incentives are removed). However, habit formation is primarily driven by an individual's internal automatic behavior with minimal involvement of external factors, which makes it difficult to disentangle from conscious learning behaviors. Therefore, identifying habit formation when learning is present would require modeling mental processes and using behavioral experiments to assess automaticity, which is beyond the scope of our data and study focus. We propose this as a potential avenue for future research.

6.3. Empirical Tests of Learning Mechanism

Now we test the existence of the learning mechanism in the Algorithm group based on the testing rule derived previously. First, for each sales worker i, we calculate $ConversionDiff_i$, which is defined as the change in the conversion rate of all bookings made by worker i from the pre-intervention period to the intervention period. For all bookings made by the Algorithm group before and after the experiment, we conduct an OLS regression analysis with the following specification:

$$UseAlgorithm_{ijt} = \theta_0 + \theta_1 ConversionDiff_i \times AfterExp_t + \tau_t + \mu_i + X_j + \epsilon_{ijt}, \tag{11}$$

Table 5 Learning: Whether Change in Experienced Conversion Rate Affects Change in Algorithm Usage

	Dependent variable: *UseAlgorithm*	
	Algorithm group	Control group
	(1)	(2)
ConversionDiff × *AfterExp*	4.3893***	0.2767
	(1.4378)	(1.5350)
Date Fixed Effects	Yes	Yes
Worker Fixed Effects	Yes	Yes
Lead Characteristics	Yes	Yes
Observations	33,336	36,700

Note: This table reports the estimation results of Equation (11) using the bookings made before and after the experiment by the Algorithm group (Column (1)) and the Control group (Column (2)), respectively. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

where θ_1 captures the impact of the change in workers' conversion rate (during-intervention compared to pre-intervention) on the change in workers' algorithm usage (post-intervention compared to pre-intervention).[22] We also replicate this analysis for bookings made by the Control group.

The results are presented in Table 5. Column (1) shows a significantly positive estimate of θ_1, suggesting that workers who have experienced a more positive change in the conversion rate during the experiment tend to increase their algorithm usage more after the experiment. This finding indicates the existence of learning in the Algorithm group according to our testing rule. The estimate of θ_1 is not statistically significant in Column (2), confirming that workers in the Control group do not update their beliefs about the algorithm during the experiment.

To visually illustrate these results, we plot in Figure 4 the change in each worker's average algorithm usage ratio from the pre-intervention period to the post-intervention period, denoted as $\Delta AlgorithmUsageRatio$, against her change in experienced conversion rate *ConversionDiff*, categorized by the Algorithm and Control groups.[23] Consistent with Table 5, Figure 4 shows a significantly positive correlation between the change in algorithm usage and the change in experienced conversion rate for workers in the Algorithm group. In contrast, no such correlation is observed in the Control group. These results again confirm that workers in the Algorithm group update their beliefs about the algorithm during the experiment, whereas workers in the Control group do not.

To further validate our finding of the existence of learning in the Algorithm Group and also to investigate whether workers update their beliefs positively or negatively, we examine whether and how the quality of leads for whom workers choose to use the algorithm changes after the experiment. If workers' beliefs about the algorithm become more positive, they should allocate more high-quality leads to

[22] Specification (11) is not intended as a causal identification, given the potential endogeneity in $ConversionDiff_i$. We conduct this test to provide supportive evidence for the learning mechanism by presenting empirical findings consistent with our theoretical implications in Section 6. Establishing causal identification for Specification (11) would ideally require an exogenous shock or natural experiment that varies algorithm exposure independently of worker characteristics, which falls outside the scope of our study.

[23] Table 15 in Appendix A presents the summary statistics of *ConversionDiff* and $\Delta AlgorithmUsageRatio$ in the Algorithm and Control groups.

Figure 4 **Change in Algorithm Usage Ratio Against Change in Experienced Conversion Rate**

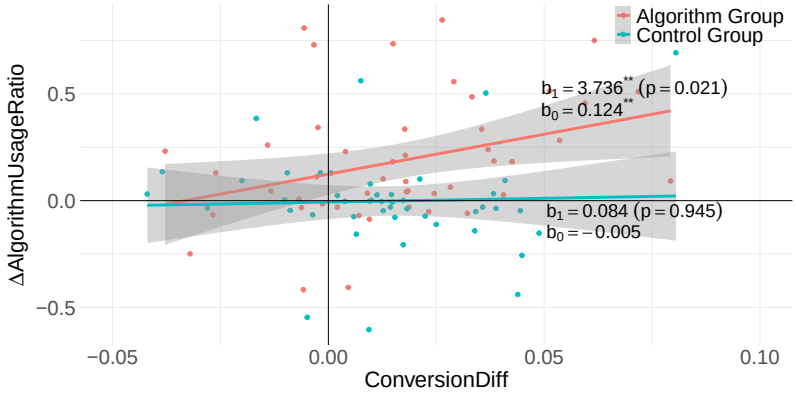


Table 6 **Post-Intervention Change in Quality of Leads and Conversion Rate of algorithmic bookings**

	Dependent variable: *QualityLead*		Dependent variable: *Purchase*	
	(1)	(2)	(3)	(4)
AlgorithmGroup× *AfterExp*	0.1128**	0.0803**	0.0175*	0.0171**
	(0.0566)	(0.0397)	(0.0097)	(0.0078)
AlgorithmGroup	-0.0520		-0.0004	
	(0.0502)		(0.0071)	
AfterExp	0.0321		0.0085	
	(0.0421)		(0.0078)	
Date Fixed Effects	No	Yes	No	Yes
Worker Fixed Effects	No	Yes	No	Yes
Observations	16,301	16,301	16,301	16,301

Note: This table reports the estimated change in quality of leads (Columns (1) and (2)) and conversion rate (Columns (3) and (4)) of algorithmic bookings made by the Algorithm group after the experiment compared to before the experiment. Columns (1) and (3) control for no fixed effects, and Columns (2) and (4) control for the date and worker fixed effects. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

algorithmic bookings. If their beliefs do not change, the quality of leads that they allocate to algorithmic bookings should not change significantly after the experiment. Thus, we can test whether and how workers' beliefs about the algorithm change in the Algorithm group by comparing the quality of leads for whom they use the algorithm before and after the experiment, using the Control group as the benchmark. Specifically, for all bookings where workers use the algorithm before and after the experiment in the Algorithm and Control groups, we perform OLS regressions with the following specifications:

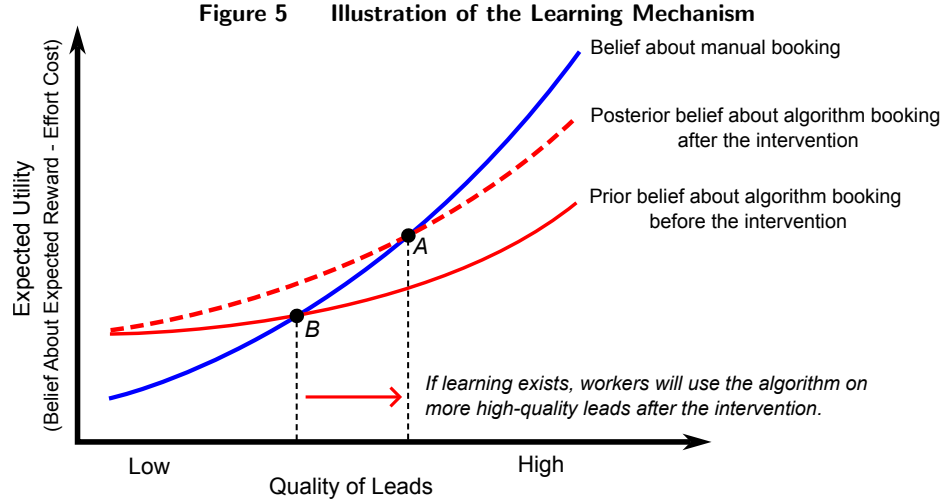
$$QualityLead_{ijt} = \eta_0 + \eta_1 AlgorithmGroup_i \times AfterExp_t + \eta_2 AlgorithmGroup_i + \eta_3 AfterExp_t + \epsilon_{ijt}, \quad (12)$$

$$QualityLead_{ijt} = \eta_0 + \eta_1 AlgorithmGroup_i \times AfterExp_t + \tau_t + \mu_i + \epsilon_{ijt}. \quad (13)$$

As shown in Columns (1) and (2) of Table 6, the coefficients of $AlgorithmGroup \times AfterExp$ are significantly positive. This suggests that compared to the Control group, workers in the Algorithm group choose to use the algorithm on more high-quality leads after the experiment.

We also conduct the same set of regressions as in Specifications (12) and (13) using $Purchase_{ijt}$ as the dependent variable and present the results in Columns (3) and (4) of Table 6. The coefficients of $AlgorithmGroup \times AfterExp$ are also significantly positive, suggesting that workers in the Algorithm group

Figure 5 Illustration of the Learning Mechanism



have a significantly larger increase in the conversion rate of algorithmic bookings after the experiment, compared to the Control group. This result further supports our finding that workers in the Algorithm group tend to use the algorithm on more high-quality leads after the experiment. These findings confirm our hypothesis in Section 6.1 that workers in the Algorithm group update their beliefs about the algorithm positively, while those in the Control group do not.[24]

We next provide an explanation, drawing on our theoretical framework in Section 6.2, for how learning through forced intervention results in improved quality of leads assigned to the algorithm and increased conversion rate of algorithmic bookings after the intervention. We illustrate our explanations with Figure 5, which depicts a worker's expected utility from a lead (belief about the expected reward from conversion minus the effort cost) against the lead quality. Note that the expected utility increases with the quality of leads given that higher-quality leads generally yield higher conversions, which dominate the utility values. We first explain why workers may selectively adopt the algorithm for more low-quality leads than high-quality ones before the experiment. As shown in Figure 5, this preference arises because the perceived utility gain of algorithmic bookings—relative to manual bookings—is lower for high-quality leads than low-quality ones. On the one hand, workers might expect the algorithm to yield smaller improvements in conversion—or higher risk of conversion loss—for high-quality leads. This is because they held negatively biased beliefs about the algorithm performance and thus perceived limited potential gains of using the algorithm when conversions from the high-quality leads are already strong. On the other hand, the perceived cost advantage of algorithmic bookings relative to manual bookings diminishes as the quality of leads increases. Specifically, algorithmic bookings yield smaller booking effort savings for high-quality leads, since workers naturally exert less effort in manual bookings for high-quality leads than low-quality ones to achieve conversions. However, algorithmic bookings may incur higher decision effort for high-quality leads, as workers may find it more difficult to evaluate whether the algorithm will outperform

[24] The conversion rates of algorithmic bookings made by the Control group before and during the experiment do not differ significantly (the estimated difference is 0.92 percentage points with p-value $= 0.124$). Thus, their negatively biased beliefs about the algorithm remain during and after the experiment.

humans in these cases. Consequently, workers perceive a lower utility advantage of algorithmic bookings for high-quality leads than low-quality ones, and thus they are more willing to use the algorithm for low-quality leads upon its initial implementation.

We now explain why, if learning exists during the forced intervention, workers in the Algorithm Group assign leads of higher quality to the algorithm and achieve improved conversion rates of algorithmic bookings after the intervention. As Figure 5 demonstrates, with negatively biased beliefs about the algorithm performance before the intervention (the solid red line), workers selectively adopt the algorithm on low-quality leads—below the threshold at point B—where they believe adopting the algorithm will increase their expected utility. The forced intervention exposes workers to the algorithm's real performance across all lead qualities. This experience shifts their beliefs about the algorithm performance upward and thereby increases the perceived utility of algorithmic bookings (the dashed red line).[25] In particular, the increase is more pronounced for high-quality leads where the algorithm was used less and the beliefs about its performance were more negatively biased. As a result, workers raise their expected gain from using the algorithm for high-quality leads and thus expand their algorithm usage to more high-quality leads (up to point A) after the intervention. Furthermore, more high-quality leads assigned to the algorithm also drives up the post-intervention conversion rates for algorithmic bookings, considering that high-quality leads inherently have higher conversion rates than low-quality ones.

7. Concluding Remarks

In this paper, we collaborate with an online education company which provided its workers with an algorithm to help them select trial class teachers for leads. We document that workers underutilized the algorithm during its cold-start stage. They tended to use the algorithm more for low-quality leads, and as a result, they observed a lower conversion rate of algorithmic bookings compared to manual bookings. The company conducted a randomized field experiment involving 171 workers that were assigned to three groups: the Algorithm group, where workers were forced to use the algorithm; the Manual group, where workers were forced to select teachers manually; and the Control group, where workers could freely choose whether to use the algorithm or not (as before the experiment). The comparison between the Algorithm group and the Control group during the experiment shows that algorithmic bookings have a conversion rate similar to that of manual bookings, confirming that the lower conversion rate of algorithmic bookings before the experiment was due to selective algorithm adoption for low-quality leads.

Using this field experiment, we study the impact of forced intervention on the persistent behavioral change in algorithm adoption among workers during the cold-start stage of algorithms. We find that forcing workers to use the algorithm for a period of time significantly increases their algorithm usage after the experiment. We further investigate the underlying mechanisms. We build a theoretical model and derive the testing rule to identify learning with the possible existence of habit formation. Our empirical

[25] Note that though algorithmic and manual bookings have comparable conversion rates in our case, their expected utilities for the same lead quality still differ due to differences in their effort costs.

evidence suggests that learning exists in the Algorithm group: forcing workers to use the algorithm for a period of time makes them learn the unbiased performance of the algorithm and improves their beliefs about the algorithm performance. As a result, it facilitates their persistent algorithm adoption, particularly for the high-quality leads.

Our results have several important practical implications. First, we show that forced intervention can significantly increase workers' post-intervention algorithm usage. This result implies that firms can utilize strong extrinsic interventions to improve workers' algorithm usage for a certain period, and they will have a long-term reinforcing impact on workers' continued algorithm usage. Second, we reveal that workers may hold negatively biased beliefs against new algorithms, leading to their underutilization of the algorithms. In particular, workers may selectively use the algorithms on low-quality cases. This suggests that firms aiming to promote algorithm systems should employ extrinsic interventions or provide educational programs to help workers learn the algorithms' real performance and form unbiased beliefs about their capabilities, so that workers become more willing to use them. While companies should carefully evaluate the risks, proper scale and timeline of algorithm implementation, it can still be valuable to promote algorithm adoption during its cold-start stage when these factors are manageable. Broader algorithm adoption can help break the cycle of biased beliefs and selective algorithm use and improve the diversity and representativeness of collected data, which can facilitate more effective algorithm training. This will be beneficial for the algorithm development and eventually the operational performance of a company in the long run. Finally, although our study focuses on algorithm aversion and algorithm usage among workers, our findings and their implications can potentially be generalized to the adoption of information technologies in more general contexts. People's resistance to new technologies is prevalent (e.g. Zwick 2002, Desmet and Parente 2014), and an important reason for the resistance to new technologies is that people have little information or knowledge about the benefits of these technologies (Joseph 2010). Our findings imply that extrinsic interventions (e.g. forced interventions) can be useful for helping people learn the benefits of new technologies and reducing the aversion to them.

Our paper also has several limitations given the limited context and treatments of our field experiment. These limitations, however, may also provide valuable opportunities for future research. First, our empirical results are derived from a sales setting with a specific algorithm. We encourage further studies on how workers' attitudes towards algorithms and the impact of extrinsic interventions vary in other settings or other human-algorithm collaboration formats. Second, we investigate just one type of extrinsic intervention—possibly the most stringent type—on algorithm aversion. Future research could examine how different types of extrinsic interventions influence humans' algorithm usage in the short term and long term and provide more guidance on the optimal interventions to reduce algorithm aversion. Third, in our empirical setting, workers cannot override algorithmic decisions; their discretion is limited to selecting which leads to assign to the algorithm. However, both human adoption of algorithms and choice to adhere to algorithmic recommendations remain essential in many real-world applications,

particularly for complex cases. An intriguing avenue for future research is to explore human-algorithm collaboration where workers determine both when to deploy the algorithm and whether to accept its recommendations. Fourth, we do not identify whether habit formation exists given the presence of learning. Future research could further explore how to distinguish the two mechanisms when both exist.

References

Acland D, Levy MR (2015) Naiveté, projection bias, and habit formation in gym attendance. *Management Science* 61(1):146–160.

Adjerid I, Ayvaci MU, Özer Ö (2023) Value of algorithm-enabled process innovation: The case of sepsis. *Manufacturing & Service Operations Management* .

Allcott H, Rogers T (2014) The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review* 104(10):3003–37.

Bai B, Dai H, Zhang DJ, Zhang F, Hu H (2022) The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *Manufacturing & Service Operations Management* 24(6):3060–3078.

Balakrishnan M, Ferreira KJ, Tong J (2025) Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science* .

Barrera-Osorio F, Bertrand M, Linden LL, Perez-Calle F (2011) Improving the design of conditional transfer programs: Evidence from a randomized education experiment in colombia. *American Economic Journal: Applied Economics* 3(2):167–95.

Becker GS (1998) Habits, addictions, and traditions. *Econometric Society Monographs* 29:123–141.

Becker GS, Murphy KM (1988) A theory of rational addiction. *Journal of Political Economy* 96(4):675–700.

Boyacı T, Canyakmaz C, de Véricourt F (2023) Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science* .

Buell RW, Cai W, Sandino T (2022) Learning or playing? The effect of gamified training on performance. *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (19-101).

Caro F, de Tejada Cuenca AS (2023) Believing in analytics: Managers' adherence to price recommendations from a dss. *Manufacturing & Service Operations Management* 25(2):524–542.

Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. *Journal of Marketing Research* 56(5):809–825.

Charness G, Gneezy U (2009) Incentives to exercise. *Econometrica* 77(3):909–931.

Chen G, Li X, Sun C, Wang H (2023) Learning to make adherence-aware advice. *arXiv preprint arXiv:2310.00817* .

Cui R, Li M, Zhang S (2022) AI and procurement. *Manufacturing & Service Operations Management* 24(2):691–706.

Dai T, Singh S (2021) Artificial intelligence on call: The physician's decision of whether to use ai in clinical practice. *Available at SSRN* .

Dargnies MP, Hakimov R, Kübler D (2024) Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science* .

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

Desmet K, Parente SL (2014) Resistance to technology adoption: The rise and decline of guilds. *Review of Economic Dynamics* 17(3):437–458.

Dietvorst BJ, Bharti S (2020) People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science* 31(10):1302–1314.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114.

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

Domingos P (2012) A few useful things to know about machine learning. *Communications of the ACM* 55(10):78–87.

Fisher M, Gallino S, Netessine S (2021) Does online training work in retail? *Manufacturing & Service Operations Management* 23(4):876–894.

Giné X, Karlan D, Zinman J (2010) Put your money where your butt is: a commitment contract for smoking cessation. *American Economic Journal: Applied Economics* 2(4):213–35.

Grand-Clément J, Pauphilet J (2024) The best decisions are not the best advice: Making adherence-aware recommendations. *Management Science* .

Gregory RW, Henfridsson O, Kaganer E, Kyriakou H (2021) The role of artificial intelligence and data network effects for creating user value. *Academy of management review* 46(3):534–551.

Gunaratne J, Zalmanson L, Nov O (2018) The persuasive power of algorithmic and crowdsourced advice. *Journal of Management Information Systems* 35(4):1092–1120.

Haftor DM, Costa-Climent R (2023) A pathway to bypassing market entry barriers from data network effects: A case study of a startup's use of machine learning. *Journal of Business Research* 158:113619.

Han BR, Li M, Zhang Y, Li P (2024) Value of autonomous last-mile delivery: Evidence from alibaba. *Available at SSRN* .

Hospedales T, Antoniou A, Micaelli P, Storkey A (2021) Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44(9):5149–5169.

Ibanez MR, Clark JR, Huckman RS, Staats BR (2018) Discretionary task ordering: Queue management in radiological services. *Management Science* 64(9):4389–4407.

Ibanez MR, Toffel MW (2020) How scheduling can bias quality assessment: Evidence from food-safety inspections. *Management Science* 66(6):2396–2416.

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

Jiang ZZ, Kong G, Zhang Y (2021) Making the most of your regret: Workers' relocation decisions in on-demand platforms. *Manufacturing & Service Operations Management* 23(3):695–713.

Joseph RC (2010) Individual resistance to IT innovations. *Communications of the ACM* 53(4):144–146.

Karlinsky-Shichor Y, Netzer O (2023) Automating the B2B salesperson pricing decisions: A human-machine hybrid approach. *Marketing Science* .

Kawaguchi K (2021) When will workers follow an algorithm? A field experiment with a retail business. *Management Science* 67(3):1670–1695.

Kesavan S, Kushwaha T (2020) Field experiment on the profit implications of merchants' discretionary power to override data-driven decision-making tools. *Management Science* 66(11):5182–5190.

Kirshner SN (2024) Psychological distance and algorithm aversion: Congruency and advisor confidence. *Service Science* .

Knight B, Mitrofanov D, Netessine S (2022) The impact of AI technology on the productivity of gig economy workers. *Available at SSRN 4372368* Working Paper.

Kwon C, Raman A, Moreno A (2023) The impact of input inaccuracy on leveraging AI tools: Evidence from algorithmic labor scheduling. *Available at SSRN* Working Paper.

Lehmann CA, Haubitz CB, Fügener A, Thonemann UW (2022) The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management* 31(9):3419–3434.

Levine SS, Jain D (2023) How network effects make ai smarter. *Harvard Business Review*, URL `https://hbr.org/2023/03/how-network-effects-make-ai-smarter`.

Li Y, Lu LX, Lu SF, Chen J (2020) The value of health IT interoperability: Evidence from interhospital transfer of heart attack patients. *Available at SSRN 3557010* .

Liang X, Wang YI, Li J (2023) Examining behavioral biases in discretionary pricing: Evidence from field and lab experiments Working paper.

Lin W, Kim SH, Tong J (2023) What drives algorithm use? An empirical analysis of algorithm use in type 1 diabetes self-management Working Paper.

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

Liu S, He L, Max Shen ZJ (2021) On-time last-mile delivery: Order assignment with travel-time predictors. *Management Science* 67(7):4095–4119.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90–103.

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4):629–650.

Lu Y, Wang Y, Chen Y, Xiong Y (2023) The role of data-based intelligence and experience on time efficiency of taxi drivers: An empirical investigation using large-scale sensor data. *Production and Operations Management* 32(11):3665–3682.

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

McElheran K, Yang MJ, Kroff Z, Brynjolfsson E (2025) The rise of industrial AI in america: Microfoundations of the productivity j-curve(s). Working Paper CES-25-27, Center for Economic Studies, U.S. Census Bureau,

URL https://www2.census.gov/library/working-papers/2025/adrm/ces/CES-WP-25-27.pdf.

Milkman KL, Minson JA, Volpp KG (2014) Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science* 60(2):283–299.

Newman DT, Fast NJ, Harmon DJ (2020) When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160:149–167.

Pierce L, Snow DC, McAfee A (2015) Cleaning house: The impact of information technology monitoring on employee theft and productivity. *Management Science* 61(10):2299–2319.

Reich T, Kaju A, Maglio SJ (2023) How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology* 33(2):285–302.

Reis L, Maier C, Mattke J, Creutzenberg M, Weitzel T (2020) Addressing user resistance would have prevented a healthcare AI project failure. *MIS Quarterly Executive* 19(4).

Royer H, Stehr M, Sydnor J (2015) Incentives, commitments, and habit formation in exercise: Evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics* 7(3):51–84.

Schein AI, Popescul A, Ungar LH, Pennock DM (2002) Methods and metrics for cold-start recommendations. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 253–260.

Schilbach F (2019) Alcohol and self-control: A field experiment in india. *American Economic Review* 109(4):1290–1322.

Senoner J, Netland T, Feuerriegel S (2022) Using explainable artificial intelligence to improve process quality: Evidence from semiconductor manufacturing. *Management Science* 68(8):5704–5723.

Song H, Tucker AL, Murrell KL, Vinson DR (2018) Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. *Management Science* 64(6):2628–2649.

Staats BR, Dai H, Hofmann D, Milkman KL (2017) Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. *Management Science* 63(5):1563–1585.

Stamatopoulos I, Bassamboo A, Moreno A (2021) The effects of menu costs on retail performance: Evidence from adoption of the electronic shelf label technology. *Management Science* 67(1):242–256.

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

Tan TF, Netessine S (2019) When you work with a superman, will you also fly? An empirical study of the impact of coworkers on performance. *Management Science* 65(8):3495–3517.

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

Turel O, Kalhan S (2023) Prejudiced against the machine? implicit associations and the transience of algorithm aversion. *Mis Quarterly* 47(4):1369–1394.

Vaccaro M, Almaatouq A, Malone TW (2024) When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour* 8:2293–2303, URL https://doi.org/10.1038/s41562-024-02024-1.

Van Donselaar KH, Gaur V, Van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Science* 56(5):766–784.

Volpp KG, Troxel AB, Pauly MV, Glick HA, Puig A, Asch DA, Galvin R, Zhu J, Wan F, DeGuzman J, et al. (2009) A randomized, controlled trial of financial incentives for smoking cessation. *N Engl J Med* 360:699–709.

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

Wang Y, Choudhary V, Yin S (2023b) Product design enhancement for fashion retailing. *Service Science* 15(3):157–171.

Xu Y, Dai H, Yan W (2023a) Identity disclosure and anthropomorphism in voice chatbot design: A field experiment Working Paper.

Xu Y, Lu B, Ghose A, Dai H, Zhou W (2023b) The interplay of earnings, ratings, and penalties on sharing platforms: An empirical investigation. *Management Science* 69(10):6128–6146.

Xu Y, Tan TF, Netessine S (2022) The impact of workload on operational risk: Evidence from a commercial bank. *Management Science* 68(4):2668–2693.

Ye Z, Zhang DJ, Zhang H, Zhang R, Chen X, Xu Z (2023) Cold start to improve market thickness on online advertising platforms: Data-driven algorithms and field experiments. *Management Science* 69(7):3838–3860.

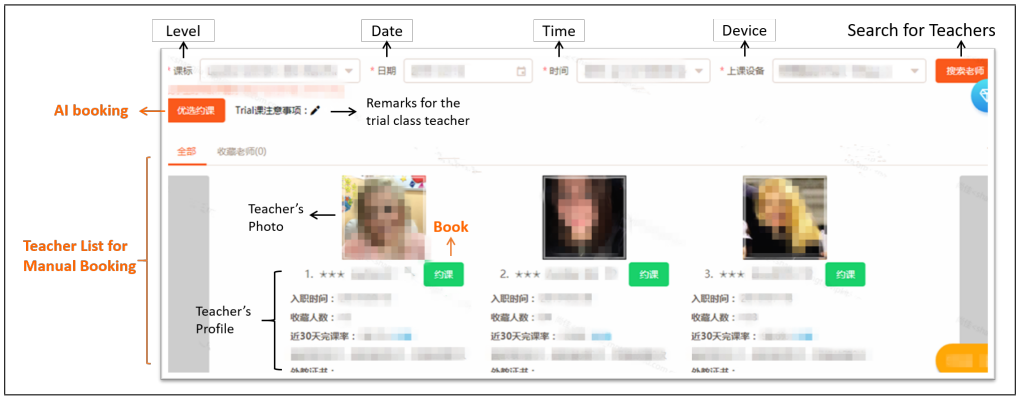
You S, Yang CL, Li X (2022) Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems* 39(2):336–365.

Zwick T (2002) Employee resistance against innovations. *International Journal of Manpower* .

Appendix A: Figures and Tables

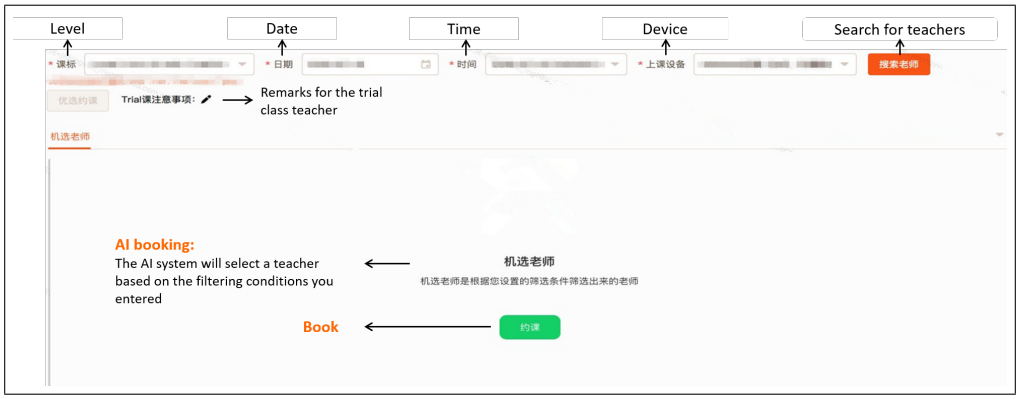
Figure 6 User Interface of Sales Workers

(a) Control group: Both manual booking and algorithmic booking are available



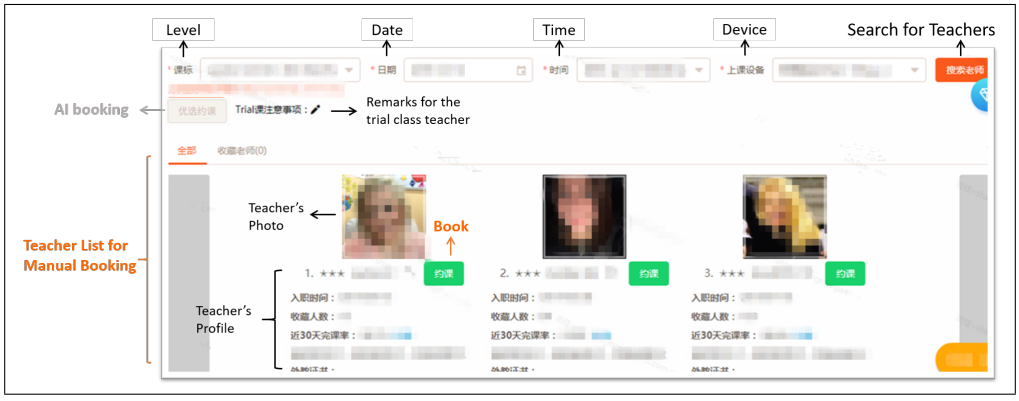
After inputting the filtering conditions, workers click the orange button "Search for teachers." Then a list of available teachers will be displayed below, and workers can manually book a teacher from the list by clicking the green button "Book," or they can click the orange button "algorithmic booking" to let the algorithm book a teacher for the lead.

(b) Algorithm group: Only algorithmic booking is available



After inputting the filtering conditions and clicking "Search for Teachers," there will be no list of teachers displayed. The workers can only click the green button "Book" to let the algorithm book a teacher for the lead.

(c) Manual Group: Only manual booking is available



The button "algorithmic booking" is disabled. Workers can only book manually.

Table 7 Definitions of Variables

Panel A: Booking-level Variables	
UseAlgorithm	Binary. Whether the booking is made using the algorithm (= 1) or manually (= 0).
Purchase	Binary. Whether the booking converts to a purchase within 17 days (= 1) or not (= 0).
BookDate	Categorical. The date on which the booking is made.
DuringExp	Binary. Whether the booking is made during the experiment (= 1) or not (= 0).
AfterExp	Binary. Whether the booking is made after the experiment (= 1) or not (= 0).
StudentAge	Categorical. Student age, ranging from 4 to 15.
QualityLead	Binary. Whether the lead is high-quality (= 1) or not (= 0), defined by the company.
Call	Binary. Whether the booking is made through a phone call (= 1) or WeChat (= 0).
Panel B: Worker-level Variables	
AlgorithmGroup	Binary. Whether the worker is assigned to the Algorithm group (= 1) or not (= 0).
Seniority	Categorical. Seniority of the worker, defined by the company according to her tenure. *Seniority* = 1 if the worker has been working in the company for less than 3 months, *Seniority* = 2 if between 3 and 6 months, *Seniority* = 3 if between 6 and 12 months, and *Seniority* = 4 if more than 12 months.
ConversionDiff	Numerical. Change in the worker's experienced conversion rate from the pre-intervention period to the during-intervention period.
Δ*AlgorithmUsageRatio*	Numerical. Change in the worker's average algorithm usage ratio from the pre-intervention period to the post-intervention period.
AlgoUsageRatioPreExp	Numerical. Average algorithm usage ratio of the worker before the experiment.
ConversionAlgoPreExp	Numercial. Conversion rate of algorithmic bookings by the worker before the experiment.
ConversionManPreExp	Numercial. Conversion rate of manual bookings by the worker before the experiment.
ConversionDiffAlgo	Numerical. Change in the conversion rate of algorithmic bookings made by the worker from the pre-intervention period to the during-intervention period.
ConversionDiffMan	Numerical. Change in the conversion rate of manual bookings made by the worker from the pre-intervention period to the during-intervention period.
Panel C: Worker-day-level Variables	
NumBooking	Numerical. Number of bookings made by the worker during the day.
AlgoUsageRatioPrevD	Numerical. Average algorithm usage ratio of the worker on the previous day.
AlgoUsageRatioPrevW	Numerical. Average algorithm usage ratio of the worker in the previous week preceding the day.
ConversionAlgoPrevD	Numercial. Conversion rate of algorithmic bookings by the worker on the previous day.
ConversionAlgoPrevW	Numercial. Conversion rate of algorithmic bookings by the worker in the previous week preceding the day.
ConversionManPrevD	Numercial. Conversion rate of manual bookings by the worker on the previous day.
ConversionManPrevW	Numercial. Conversion rate of manual bookings by the worker in the previous week preceding the day.

Table 8 Distribution of Sales Workers Across Sales Areas and Offices

Sales Area	Sales Office	Treatment Group			Total
		Algorithm Group	Manual Group	Control Group	
Beijing	First Office	7	5	8	20
		(11.86)	(9.43)	(13.56)	(11.70)
	Second Office	4	0	3	7
		(6.78)	(0.00)	(5.08)	(4.09)
	Third Office	7	4	4	15
		(11.86)	(7.55)	(6.78)	(8.77)
	Fourth Office	2	3	5	10
		(3.39)	(5.66)	(8.47)	(5.85)
	Sixth Office	2	3	2	7
		(3.39)	(5.66)	(3.39)	(4.09)
	Seventh Office	5	2	6	13
		(8.47)	(3.77)	(10.17)	(7.60)
	Ninth Office	1	5	1	7
		(1.69)	(9.43)	(1.69)	(4.09)
	Experienced Office	6	8	3	17
		(10.17)	(15.09)	(5.08)	(9.94)
Shanghai	Fifth Office	3	3	6	12
		(5.08)	(5.66)	(10.17)	(7.02)
	Tenth Office	11	9	12	32
		(18.64)	(16.98)	(20.34)	(18.71)
Chengdu	Eleventh Office	1	0	1	2
		(1.69)	(0.00)	(1.69)	(1.17)
	Twelfth Office	5	4	2	11
		(8.47)	(7.55)	(3.39)	(6.43)
Shenzhen	Fifteenth Office	5	7	6	18
		(8.47)	(13.21)	(10.17)	(10.53)
	Total	59	53	59	171
		(100.00)	(100.00)	(100.00)	(100.00)

Note: This table reports the number of workers in each sales office allocated to each treatment group. We report the percentage that it accounts for the total number of workers in the corresponding treatment group in parentheses. The Fisher's exact test across sales areas has p-value$= 0.882$, and the Fisher's exact test across sales offices has p-value$= 0.697$. The test results indicate that the three treatment groups do not have significantly different distributions of workers across sales areas or sales offices.

Table 9 Balance Check: Summary Statistics Before the Experiment (Three Groups)

	Each Group: Mean (Std. Dev.)			Between-group Comparison: p-value		
	Algorithm	Manual	Control	Algorithm vs. Control	Manual vs. Control	Algorithm vs. Manual
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Booking-level Variables						
UseAlgorithm	0.213	0.181	0.215	0.948	0.391	0.403
	(0.409)	(0.385)	(0.411)			
Purchase	+0.0015	−0.0024	–	0.743	0.616	0.419
	–	–	–			
StudentAge	8.140	8.139	8.143	0.963	0.952	0.993
	(2.667)	(2.634)	(2.644)			
QualityLead	0.519	0.502	0.548	0.466	0.255	0.676
	(0.500)	(0.500)	(0.498)			
Call	0.829	0.848	0.824	0.755	0.084*	0.117
	(0.377)	(0.359)	(0.381)			
Observations	27,696	26,365	29,266			
Panel B: Worker-level Variables						
Seniority	2.559	2.396	2.492	0.765	0.682	0.486
	(1.236)	(1.230)	(1.223)			
Observations	59	53	59			
Panel C: Worker-day-level Variables						
NumBooking	6.189	6.638	6.602	0.134	0.911	0.166
	(3.158)	(3.488)	(3.254)			
NumPurchase	−0.0069	−0.0146	–	0.835	0.673	0.825
	(−)	(−)	(−)			
Observations	4,475	3,972	4,433			

Note: This table presents the summary statistics and balance check results of the key variables before the experiment in each group. Columns (1)-(3) report the mean and standard deviation of each variable in each group before the experiment, with standard deviations given in parentheses. Columns (4)-(6) report the p-values of the three t-tests that compare the means of the variables in each pair of the groups. Due to the NDA, we cannot report the means and standard deviations of *Purchase* and *NumPurchase*. Instead, we report the difference in the means of *Purchase* and *NumPurchase* among the three groups, using the Control group as the benchmark. Standard errors are clustered at the worker level for the t-tests in Panels A and C. *p<0.1; **p<0.05; ***p<0.01.

Table 10 Balance Check: Summary Statistics During the Experiment (Three Groups)

	Each Group: Mean (Std. Dev.)			Between-group Comparison: p-value		
	Algorithm	Manual	Control	Algorithm vs. Control	Manual vs. Control	Algorithm vs. Manual
	(1)	(2)	(3)	(4)	(5)	(6)
StudentAge	8.057	7.909	7.939	0.185	0.739	0.102
	(2.708)	(2.636)	(2.614)			
QualityLead	0.571	0.535	0.563	0.835	0.497	0.382
	(0.495)	(0.499)	(0.496)			
Call	0.830	0.831	0.823	0.681	0.630	0.956
	(0.375)	(0.375)	(0.382)			
Observations	4,118	4,534	5,051			

Note: This table presents the summary statistics of the lead characteristics during the experiment in each group. Columns (1)-(3) report the mean and standard deviation of each variable in each group during the experiment, with standard deviations given in parentheses. Columns (4)-(6) report the p-values of the three t-tests that compare the means of the variables in each pair of the groups. Standard errors are clustered at the worker level. *p<0.1, **p<0.05, ***p<0.01.

Table 11 Balance Check: Summary Statistics After the Experiment (Three Groups)

	Each Group: Mean (Std. Dev.)			Between-group Comparison: p-value		
	Algorithm	Manual	Control	Algorithm vs. Control	Manual vs. Control	Algorithm vs. Manual
	(1)	(2)	(3)	(4)	(5)	(6)
StudentAge	8.487	8.356	8.453	0.711	0.269	0.124
	(2.751)	(2.683)	(2.691)			
QualityLead	0.629	0.570	0.608	0.601	0.331	0.155
	(0.483)	(0.495)	(0.488)			
Call	0.808	0.822	0.814	0.732	0.675	0.422
	(0.394)	(0.383)	(0.389)			
Observations	7,005	7,493	7,677			

Note: This table presents the summary statistics of the lead characteristics after the experiment in each group. Columns (1)-(3) report the mean and standard deviation of each variable in each group after the experiment, with standard deviations given in parentheses. Columns (4)-(6) report the p-values of the three t-tests that compare the means of the variables in each pair of the groups. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

Table 12 Number of Workers Before and After the Experiment (Three Groups)

	Algorithm Group	Manual Group	Control Group	Total
Experiment Participant Count (May 30, 2019)	59	53	59	171
Post-experiment Participant Count (June 19, 2019)	48	46	47	141
Probability of Staying After the Experiment	0.814	0.868	0.797	0.825
	Algorithm vs. Control	Manual vs. Control	Algorithm vs. Manual	
p-value of Chi-squared Test	0.816	0.315	0.434	

Note: This table reports the number of workers that participated in and stayed after the experiment in each group. It also reports the retention rate of each group and across all the three groups, along with the p-values of the chi-squared tests on the retention rates between each pair of the groups. *p<0.1; **p<0.05; ***p<0.01.

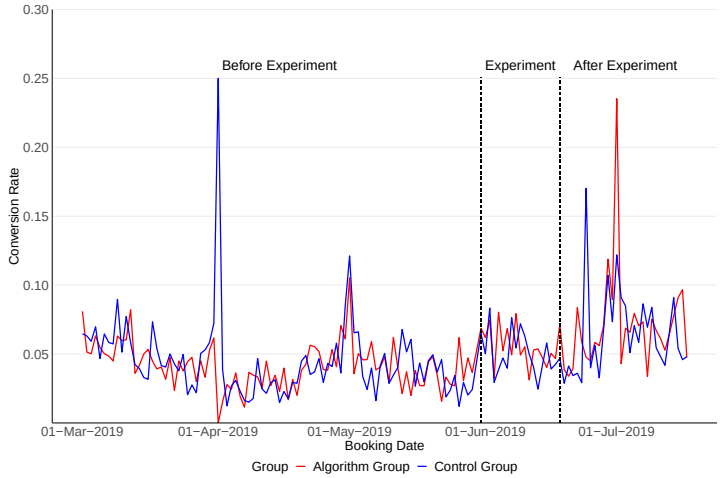
Table 13 Effects of Using Algorithm on Conversion Rate During Experiment

	Dependent variable: *Purchase*	
	(1)	(2)
AlgorithmGroup × *DuringExp*	0.0058	0.0028
	(0.0056)	(0.0051)
AlgorithmGroup	0.0015	
	(0.0046)	
DuringExp	0.0111***	
	(0.0039)	
Date Fixed Effects	No	Yes
Worker Fixed Effects	No	Yes
Lead Characteristics	No	Yes
Observations	66,131	66,131

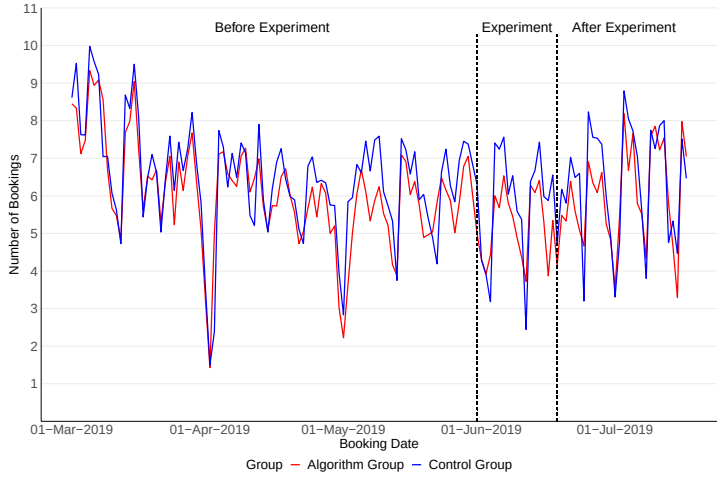
Note: This table reports the estimated treatment effects of forced algorithm use on conversion rate during the experiment in the Algorithm group compared to the Control group. We use Specifications (1) and (2) with the dependent variable being $Purchase_{ijt}$ and $AfterExp_t$ replaced by $DuringExp_t$ for estimation. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

Figure 7 Daily Average Operational Performance

(a) Daily Average Conversion Rate



(b) Daily Average Number of Bookings per Worker



(c) Daily Average Number of Converted Bookings per Worker

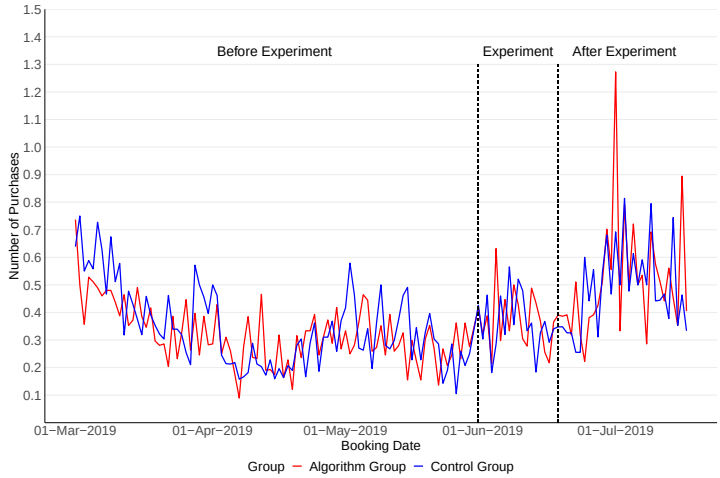


Table 14 Empirical Tests for No Learning Before the Experiment

	Dependent variable: *UseAlgorithm*			
	(1)	(2)	(3)	(4)
AlgoUsageRatioPrevD	0.3707***			
	(0.0376)			
ConversionAlgoPrevD	0.0201			
	(0.0231)			
ConversionManPrevD	0.0146			
	(0.0451)			
AlgoUsageRatioPrevW		0.6190***		
		(0.0467)		
ConversionAlgoPrevW		-0.0370		
		(0.0323)		
ConversionManPrevW		0.0319		
		(0.0759)		
AlgoUsageRatioPreExp			0.6263***	
			(0.1522)	
ConversionAlgoPreExp			-0.0719	
			(0.2876)	
ConversionManPreExp			0.9931	
			(1.0412)	
ConversionAlgoPreExp × *DuringExp*				-0.2637
				(0.3997)
ConversionManPreExp × *DuringExp*				0.7460
				(1.1884)
Date Fixed Effects	Yes	Yes	Yes	Yes
Worker Fixed Effects	Yes	Yes	No	Yes
Lead Characteristics	Yes	Yes	Yes	Yes
Observations	22,642	36,408	3,930	28,026

Note: This table reports the empirical test results for the non-existence of learning before the experiment. The definitions of all the variables are given in Table 7. Columns (1)-(2) use all bookings made by the Algorithm and Control groups before the experiment where a worker has ever used the algorithm for bookings on the previous day (Column (1)) or in the previous week preceding the day of observation (Column (2)). Columns (3) and (4) use all bookings during the experiment (Column (3)) and before and during the experiment (Column (4)) which are made by workers in the Control group that have ever used algorithm for bookings before the experiment. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

Columns (1)-(2) show that before the experiment, workers' algorithm usage is positively correlated with their past algorithm usage ratio (on the previous day or in the previous week), but is not affected by the past performance of algorithmic bookings or manual bookings (i.e., the conversion rate of algorithmic bookings or manual bookings made on the previous day or in the previous week). To address the potential concern that one day or one week may be too short for workers to learn the performance of the algorithm, we also explore whether the performance of the algorithm before the experiment can influence workers' algorithm usage during the experiment in Columns (3) and (4). We utilize the data in the Control group for this analysis, since the Control group can freely choose whether to use the algorithm for each lead during the experiment. Columns (3) and (4) show that even if we use the entire pre-experiment period to measure the performances of algorithm or manual bookings, they still do not affect workers' algorithm usage during the experiment. This further validates that there is no learning before the experiment.

Table 15 Summary Statistics of ConversionDiff and ΔAlgorithmUsageRatio

	Statistics	Algorithm group	Control Group
ConversionDiff	Mean	0.0164	0.0156
	Ste. Dev.	0.0264	0.0278
	Min	-0.0378	-0.0420
	Max	0.0793	0.1151
$\Delta AlgorithmUsageRatio$	Mean	0.186	-0.003
	Std. Dev.	0.293	0.228
	Min	-0.417	-0.604
	Max	0.846	0.693
	Observations	47	47

Note: This table reports the summary statistics of *ConversionDiff* and $\Delta AlgorithmUsageRatio$ of all sales workers who have non-empty values of these two variables in the Algorithm and Control groups. Notice that the number of observations in the Algorithm group is smaller than the number of workers who stayed after the experiment in this group, because there is one sales worker in the Algorithm group who stayed after the experiment but it happened that they made no bookings during the experiment, and therefore their *ConversionDiff* is missing.

Table 16 Learning: Empirical Tests Using Alternative Definitions of ConversionDiff

	Dependent variable: *UseAlgorithm*		
	(1)	(2)	(3)
AfterExp × *ConversionDiffAlgo*	1.1998*	0.5166	
	(0.6857)	(0.5198)	
AfterExp × *ConversionDiffMan*			0.1470
			(1.1342)
Date Fixed Effects	Yes	Yes	Yes
Worker Fixed Effects	Yes	Yes	Yes
Lead Characteristics	Yes	Yes	Yes
Observations	32,277	34,739	36,700

Note: This table reports estimation results of Specificaion (11) using alternative measures of the change in conversion rates of a worker, i.e., *ConversionDiffAlgorithm* and *ConversionDiffManual*, with definitions given in Table 7. Column (1) reports the results using the bookings made by the Algorithm group before and after the experiment. Columns (2) and (3) reports the results using the bookings made by the Control group before and after the experiment with non-empty values of *ConversionDiffAlgo* and *ConversionDiffMan*, respectively. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

Column (1) confirms that learning exists in the Algorithm group. The reason why the coefficient is smaller than that in Column (1) of Table 5 is possibly because most workers have low algorithm usage before the experiment. In other words, it is likely that workers in the Algorithm group learn the performance of the algorithm by comparing the conversion rate during the experiment to the overall conversion rate before the experiment rather than the conversion rate of algorithmic bookings before the experiment. Columns (2)-(3) again confirm that there is no learning in the Control group.

Appendix B: Additional Analyses

B.1. Additional Discussion on Workers' Exiting Behavior

Table 12 in Appendix A reports the numbers of workers that participated in the experiment and stayed after the experiment, as well as their probabilities of staying after the experiment, for both the Algorithm and Control groups. We notice that 23 workers quited their jobs during the experiment in the two groups. The probability of a worker staying after the experiment is not statistically different between the two groups, and the drop-out rate of the Control group is directionally higher than that of the Algorithm group. Therefore, the treatment (i.e., forcing workers to use the algorithm) cannot be the major driver of these workers' dropouts.

Moreover, in Columns (2) and (3) of Table 1 in Section 5.1, including worker fixed effects in the regression automatically removes the workers who quitted the jobs during the experiment and restricts the analysis to those who stayed after the experiment. If the workers who quitted during the experiment are systematically different between the two groups (even though the average quitting probabilities are similar between the two groups), the estimate of the post-intervention effect with worker fixed effects controlled for will be significantly different from that without worker fixed effects controlled for. In other words, the change in the estimation results reflects the extent to which the post-intervention effect is influenced by workers' self-selection behavior, i.e., workers in the Algorithm group who prefer using the algorithm are more likely to stay after the experiment. Table 1 shows that after adding worker fixed effects, the estimated post-intervention effect for the Algorithm group gets slightly larger but qualitatively unchanged. This indicates that the post-intervention effect on the Algorithm group is not due to workers' self-selection but rather the behavioral change of the workers who stayed, i.e., forcing workers to use the algorithm makes them more likely to continue using it after the experiment.

B.2. Post-intervention Algorithm Usage at the Individual Level

In this section, we examine the individual-level effect of the forced intervention on workers' algorithm usage ratio after the experiment. Specifically, for workers in the Algorithm and Control groups, we conduct OLS regression analysis using the following specification:

$$AvgUseAlgorithm_{it} = \lambda_0 + \lambda_1 AlgorithmGroup_i \times AfterExp_t + \lambda_2 AlgorithmGroup_i + \lambda_3 AfterExp_t + \epsilon_{it}, \tag{14}$$

$$AvgUseAlgorithm_{it} = \lambda_0 + \lambda_1 AlgorithmGroup_i \times AfterExp_t + \tau_t + \mu_i + \epsilon_{it}, \tag{15}$$

where $AvgUseAlgorithm_{it}$ denotes the average algorithm usage ratio of all bookings made by worker i on day t and all other variables are defined as in the paper. Table 17 presents the estimation results. We find that forced intervention significantly increases the post-intervention algorithm usage ratio for workers in the Algorithm group compared to the Control group – by 73.10% without fixed effects included and 83.39% with date and worker fixed effects included in our estimation. These results are consistent with our main findings in Section 5.1, i.e., forced intervention leads to a significant increase in the post-intervention algorithm usage for workers in the Algorithm group.

Table 17 Post-Intervention Effect on Individual-level Algorithm Usage

	Dependent variable: *AvgUseAlgorithm*	
	(1)	(2)
AlgorithmGroup × *AfterExp*	0.1549***	0.1767***
	(0.0559)	(0.0639)
AlgorithmGroup	−0.0023	
	(0.0382)	
AfterExp	−0.0005	
	(0.0340)	
Relative Effect Size	73.10%	83.39%
Date Fixed Effects	No	Yes
Worker Fixed Effects	No	Yes
Observations	213	213

Note: This table reports the estimated treatment effects on sales workers' individual-level algorithm usage after the intervention. Columns (1) does not control for date and worker fixed effects while Column (2) controls for these fixed effects. The relative effect size is computed based on the average algorithm usage of workers from the Algorithm group before the experiment. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01.

B.3. Heterogeneous Treatment Effects of Forced Intervention

In this section, we explore the heterogeneous treatment effects of forced intervention on post-intervention algorithm usage ratio across workers with different characteristics. Specifically we examine how the treatment effects are moderated by the following variables: the seniority of a worker ($Seniority_i$), the algorithm usage ratio of a worker before the experiment ($AlgoUsageRatioPreExp_i$), and the conversion rate of all bookings made by a worker before the experiment ($ConversionPreExp_i$). We conduct the OLS regression analyses with the following specification:

$$\begin{aligned} UseAlgorithm_{ijt} =& \alpha_0 + \alpha_1 AlgorithmGroup_i \times AfterExp_t + \alpha_2 Moderator_i \times AfterExp_t \\ &+ \alpha_3 AlgorithmGroup_i \times AfterExp_t \times Moderator_i + \tau_t + \mu_i + X_j + \epsilon_{ijt}, \end{aligned} \tag{16}$$

where $Moderator_i \in \{Seniority_i, AlgoUsageRatioPreExp_i, ConversionPreExp_i\}$ and all other variables are defined as in the paper. Table 18 reports the estimation results. We find that the coefficient for *AlgorithmGroup* × *Seniority* × *AfterExp* lacks statistical significance, showing that workers of different seniority levels do not exhibit differential changes in algorithm usage in response to the forced intervention. Similarly, the triple interaction coefficients in Columns (2) and (3) are also insignificant, indicating that the forced intervention does not produce heterogeneous effects on algorithm adoption across workers with varying pre-intervention algorithm usage ratio or conversion rate. These findings collectively demonstrate that the impact of forced intervention on post-intervention algorithm usage has no significant heterogeneity across workers with different seniority or past performance measures.

B.4. Power Analysis for the Post-intervention Operational Performance Evaluation

Given that the sample size for the individual-level analysis in Table 4 is smaller than that in other analyses of the paper, the corresponding insignificant findings require an examination of statistical power. Specifically, we conduct a Monte Carlo power analysis to examine whether our sample size is sufficient to detect meaningful treatment effects. The simulation procedure involves generating 1,000 datasets that replicate the structure characteristics of our experiment, including the number of workers, observation days, and the empirically observed within- and between-worker variance. We then analyze each simulated dataset using Specifications (4) and (5). The resulting statistical power is subsequently calculated as the proportion of these simulations in which the interaction term for the treatment effect is statistically significant at the 5% level.

Table 18 Heterogeneous Treatment Effects on Algorithm Usage Ratio

	Dependent variable: *UseAlgorithm*		
	(1)	(2)	(3)
AlgorithmGroup × *Seniority* × *AfterExp*	0.0255		
	(0.0401)		
AlgorithmGroup × *AlgoUsageRatioPreExp* × *AfterExp*		0.0479	
		(0.2972)	
AlgorithmGroup × *ConversionPreExp* × *AfterExp*			−1.0246
			(2.0516)
Seniority × *AfterExp*	−0.0214		
	(0.0239)		
AlgoUsageRatioPreExp × *AfterExp*		−0.4234**	
		(0.1773)	
ConversionPreExp × *AfterExp*			−0.4418
			(0.9512)
AlgorithmGroup × *AfterExp*	0.1314	0.1986***	0.2436
	(0.1202)	(0.0702)	(0.1006)
Date Fixed Effects	Yes	Yes	Yes
Worker Fixed Effects	Yes	Yes	Yes
Lead Characteristics	Yes	Yes	Yes
Observations	71,644	71,644	71,644

Note: This table reports the heterogeneous treatment effects on workers' post-intervention algorithm usage. Standard errors are clustered at the worker level. *p<0.1; **p<0.05; ***p<0.01

Specifically, we evaluate the statistical powers of detecting 5% and 10% effect sizes in Specifications (4) and (5) in the paper (i.e., the treatment effect λ_1 is 5% or 10% of the pre-intervention average of the Control group) given our sample size. For Specification (4), our sample size achieves a statistical power of 84.7% for detecting a 5% effect size and 100% for a 10% effect size. For Specification (5), the statistical powers of our sample size are 89.7% and 100% for detecting 5% and 10% effect sizes, respectively. These results suggest that our sample size is sufficient to detect small-to-moderate treatment effects. Thus, the insignificant interaction terms in Columns (3)-(6) of Table 4 likely reflect the treatment effects being absent or too small to be detected, rather than a result of insufficient sample size.

Appendix C: Extension of the Theoretical Model in Section 6.2

C.1. Extension to General Utility Functions

In this section, we extend the theoretical framework of Section 6.2 by allowing for more general utility functions. We demonstrate that testing rule established in Section 6.2 remains valid: If learning exists, a worker who experiences a greater positive update in her belief about algorithm performance during the intervention (relative to before the intervention) should exhibit a larger increase in her algorithm usage ratio after the intervention (relative to before).

Consider three time periods: $t=0$ (before the experiment), $t=1$ (during the experiment), and $t=2$ (after the experiment). In each period t, worker i holds a belief about algorithm performance, $v_{ti} \sim N(\mu_{ti}, \sigma_{ti}^2)$. When $t=0$ or 2, worker i chooses $a_{ti} \in [0,1]$, the probability of using the algorithm to book a trial class for a lead in period t. The worker selects a_{ti} to maximize her expected utility from serving a lead, given her belief $v_{t-1,i}$ and action $a_{t-1,i}$ in the previous period $t-1$:

$$U(a_{ti}; \mu_{t-1,i}, a_{t-1,i}) = \mathrm{E}[R(a_{ti}, v_{t-1,i}) \mid \mu_{t-1,i}] - C(a_{ti}) - H(|a_{ti} - a_{t-1,i}|). \tag{17}$$

Here, $R(a_{ti}, v_{t-1,i})$ denotes the reward from serving a lead given the algorithm usage ratio a_{ti} and belief about algorithm performance in the last period $v_{t-1,i}$. $C(a_{ti})$ is the effort cost of serving a lead given the algorithm usage ratio a_{ti}. $H(|a_{ti} - a_{t-1,i}|)$ is a "habit cost" that captures the disutility of deviating from the previous algorithm usage level. In period $t = 0$, the habit cost should be omitted, as there is no prior action. This omission does not affect our subsequent analysis, so we do not discuss this case separately. For simplicity, we drop the subscripts i and t where there is no ambiguity, and denote a_0 as the action in the previous period to distinguish it from the action a in the current period.

We impose the following assumptions on the model primitives. First, the reward function $R(a, v)$ is twice continuously differentiable in both arguments on $[0, 1] \times V$, where $V \subseteq \mathbb{R}$, and weakly concave in a. The weak concavity of R in a reflects the diminishing marginal returns from using the algorithm progressively on less promising leads. We also assume the cross-partial derivative satisfies $R_{av}(a, v) = \partial^2 R/\partial a \partial v > 0$ for all (a, v), which captures the idea that an improvement in the belief about algorithm performance raises the marginal benefit of increasing the algorithm usage level. Second, the cost function $C(a)$ is twice continuously differentiable and strictly convex in a. The convexity of $C(a)$ reflects increasing marginal effort cost as a increases: As workers expand algorithm usage on additional leads, the marginal savings in booking effort (i.e., effort spent on selecting teachers) decreases, since workers tend to automate the most time-consuming manual bookings first. However, the marginal decision effort increases, as it becomes increasingly difficult to determine on borderline cases whether relying on the algorithm or personal judgment is better. Third, the habit cost $H(x)$ is twice continuously differentiable, non-decreasing, and weakly convex in $x \geq 0$. These conditions capture the idea that the disutility of deviating from previous algorithm usage level increases with the magnitude of deviation, and the marginal cost of further deviation is non-decreasing. Finally, we assume that the marginal utility of increasing algorithm usage at the endpoints satisfy $U_a(0; \mu, a_0) > 0$ and $U_a(1; \mu, a_0) < 0$ for all μ and a_0. This reflects the situation that a worker finds it worthwhile to use the algorithm on some, but not all, leads, which is typical in the cold-start stage of algorithm development as discussed in the paper.

We now show that, for any fixed a_0, the worker's optimal action $a^*(\mu, a_0) = \arg\max_{a \in [0,1]} U(a; \mu, a_0)$ is strictly increasing in μ. First, under the stated assumptions, $R(a, v)$ is concave in a, while $C(a)$ and $H(|a - a_0|)$ are convex in a, with $C(a)$ strictly convex. Thus, $U(a; \mu, a_0)$ is strictly concave in a, and the boundary conditions ensure the existence of a unique interior maximizer $a^*(\mu, a_0) \in (0, 1)$ for every μ and a_0.

Next, we compute the cross-partial derivative of the utility function with respect to a and μ. Since C and H do not depend on μ, we have

$$U_{a\mu}(a; \mu, a_0) = \frac{\partial}{\partial a \partial \mu} \int_{-\infty}^{\infty} R(a, v) f(v; \mu) dv = \frac{\partial}{\partial \mu} \int_{-\infty}^{\infty} R_a(a, v) f(v; \mu) dv, \tag{18}$$

where $f(v; \mu)$ denotes the density function of the normal distribution $N(\mu, \sigma^2)$. By the dominated convergence theorem,

$$\frac{\partial}{\partial \mu} \int R_a(a, v) f(v; \mu) dv = \int R_a(a, v) \frac{\partial}{\partial \mu} f(v; \mu) dv. \tag{19}$$

Recall that $\frac{\partial}{\partial \mu} f(v; \mu) = \frac{v - \mu}{\sigma^2} f(v; \mu)$. Therefore,

$$U_{a\mu}(a; \mu, a_0) = \frac{1}{\sigma^2} \int R_a(a, v)(v - \mu) f(v; \mu) dv = \frac{1}{\sigma^2} \operatorname{Cov}_\mu(R_a(a, V), V). \tag{20}$$

By Stein's lemma for the normal distribution,

$$\mathrm{Cov}_\mu(R_a(a,V),V) = \sigma^2 \mathbb{E}_\mu[R_{av}(a,V)], \tag{21}$$

and thus

$$U_{a\mu}(a;\mu,a_0) = \mathbb{E}_\mu[R_{av}(a,V)] > 0, \tag{22}$$

where the positivity follows from $R_{av}(a,v) > 0$ for all (a,v).

The monotonicity of $a^*(\mu,a_0)$ in μ now follows from the implicit function theorem. The first-order condition for optimality is

$$\frac{\partial U}{\partial a}(a^*(\mu,a_0);\mu,a_0) = 0. \tag{23}$$

Differentiating both sides with respect to μ yields

$$U_{aa}(a^*(\mu,a_0);\mu,a_0) \cdot \frac{\partial a^*(\mu,a_0)}{\partial \mu} + U_{a\mu}(a^*(\mu,a_0);\mu,a_0) = 0, \tag{24}$$

and thus

$$\frac{\partial a^*(\mu,a_0)}{\partial \mu} = -\frac{U_{a\mu}(a^*(\mu,a_0);\mu,a_0)}{U_{aa}(a^*(\mu,a_0);\mu,a_0)}. \tag{25}$$

Because $U_{a\mu} > 0$ and $U_{aa} < 0$ (by strict concavity), it follows that $\frac{\partial a^*(\mu,a_0)}{\partial \mu} > 0$. Therefore, the worker's optimal algorithm usage $a^*(\mu,a_0)$ is strictly increasing in her belief about algorithm performance μ, holding the previous action a_0 fixed.

To conclude, for a worker i in the algorithm group, let $\Delta\mu_i = \mu_{1i} - \mu_{0i}$ and $\Delta a_i = a^*_{2i}(\mu_{1i},a_{1i}) - a^*_{0i}(\mu_{0i})$. Since $a^*_{2i}(\mu_{1i},a_{1i})$ is strictly increasing in μ_{1i} for given a_{1i}, it follows that, holding μ_{0i} and $a^*_{0i}(\mu_{0i})$ fixed, Δa_i is strictly increasing in $\Delta\mu_i$. This establishes that the testing rule of Section 6.2 holds under the more general utility formulation considered here.

C.2. Further Extension to Day-to-Day Variation in a Worker's Behavior

We can extend the model further by refining the temporal granularity of the model so that each period t corresponds to one calendar day in the experiment. This allows for day-to-day variation in a worker's belief about the algorithm performance $v_t \sim N(\mu_t,\sigma_t^2)$ and in her algorithm usage ratio a_t. For notational simplicity, we omit the worker index i. Let T_0, T_1, T_2 denote the sets of days corresponding to the pre-experiment, experiment, and post-experiment periods, respectively. Assume that, for each k, the sequence $\{\mu_t(\bar\mu_k) : t \in T_{k+1}\}$ is a family of random vectors indexed by $\bar\mu_k$, where $\bar\mu_k = \mathbb{E}[\frac{1}{|T_{k+1}|}\sum_{t\in T_{k+1}} \mu_t]$. That is, a worker's belief about algorithm performance in period T_{k+1} is a random vector with component-wise average being her average belief formed in the previous period T_k. We require that $\{\mu_t(\bar\mu_k) : t \in T_{k+1}\}$ is strictly stochastically increasing in $\bar\mu_k$; that is, for any $\bar\mu_k^1 < \bar\mu_k^2$, the random vector $\{\mu_t(\bar\mu_k^2) : t \in T_{k+1}\}$ strictly stochastically dominates $\{\mu_t(\bar\mu_k^1) : t \in T_{k+1}\}$.

Define $\Delta\bar\mu = \bar\mu_1 - \bar\mu_0$ and $\Delta\bar a = \bar a_2^* - \bar a_0^*$, where $\bar a_k^* = \frac{1}{|T_k|}\sum_{t\in T_k} a_t^*(\mu_{t-1},a_{t-1})$ is the average optimal algorithm usage ratio during period T_k. We aim to show that the testing rule in Section 6.2 remains valid under this extension; specifically, that $\Delta\bar a$ is strictly increasing in $\Delta\bar\mu$ almost surely.

Recall the first-order condition characterizing the optimal algorithm usage ratio $a^*(\mu, a_0)$:

$$U_a(a^*(\mu, a_0); \mu, a_0) = 0, \tag{26}$$

where U is strictly concave in a and twice continuously differentiable. Differentiating both sides with respect to a_0 yields

$$U_{aa}(a^*(\mu, a_0); \mu, a_0) \cdot \frac{\partial a^*(\mu, a_0)}{\partial a_0} + U_{aa_0}(a^*(\mu, a_0); \mu, a_0) = 0, \tag{27}$$

so

$$\frac{\partial a^*(\mu, a_0)}{\partial a_0} = -\frac{U_{aa_0}(a^*(\mu, a_0); \mu, a_0)}{U_{aa}(a^*(\mu, a_0); \mu, a_0)}. \tag{28}$$

Given that $H(x)$ is non-decreasing and weakly convex for $x \geq 0$ and enters U as a cost $-H(|a - a_0|)$, we compute:

$$U_{aa_0}(a, \mu, a_0) = H''(|a - a_0|) \geq 0. \tag{29}$$

Since $U_{aa} < 0$ (by strict concavity of U) and $U_{aa_0} \geq 0$, we have $\frac{\partial a^*}{\partial a_0} \geq 0$; that is, $a^*(\mu, a_0)$ is weakly increasing in a_0.

We next show that $\bar{a}_2^*$ strictly increases with $\bar{\mu}_1$ almost surely. Take any $\bar{\mu}_1^1 < \bar{\mu}_1^2$. By the stochastic dominance assumption and Strassen's theorem, there exists a coupling such that, $\mu_t(\bar{\mu}_1^1) < \mu_t(\bar{\mu}_1^2)$ for all $t \in T_2$ almost surely. Denote τ_2 as the first day in T_2. Fixing $a_{\tau_2 - 1}$, let $\{a_t^*(\bar{\mu}_1^1) : t \in T_2\}$ and $\{a_t^*(\bar{\mu}_1^2) : t \in T_2\}$ be the sequences of optimal actions in period T_2 corresponding to the sequences of beliefs $\{\mu_t(\bar{\mu}_1^1) : t \in T_2\}$ and $\{\mu_t(\bar{\mu}_1^2) : t \in T_2\}$, respectively.

We proceed by induction. On the first day $t = \tau_2$, since $a^*(\mu, a_0)$ is strictly increasing in μ and weakly increasing in a_0, we have

$$a_{\tau_2}^*(\mu_{\tau_2 - 1}(\bar{\mu}_1^1), a_{\tau_2 - 1}) < a_{\tau_2}^*(\mu_{\tau_2 - 1}(\bar{\mu}_1^2), a_{\tau_2 - 1}). \tag{30}$$

almost surely. Assume that for some $t > \tau_2$, $a_{t-1}^*(\bar{\mu}_1^1) < a_{t-1}^*(\bar{\mu}_1^2)$ almost surely. Then, since $a^*(\mu, a_0)$ is strictly increasing in μ and weakly increasing in a_0,

$$a_t^*(\mu_{t-1}(\bar{\mu}_1^1), a_{t-1}^*(\bar{\mu}_1^1)) < a_t^*(\mu_{t-1}(\bar{\mu}_1^2), a_{t-1}^*(\bar{\mu}_1^2)) \tag{31}$$

almost surely. By induction, this holds for all $t \in T_2$.

Therefore,

$$\bar{a}_2^*(\bar{\mu}_1^1) = \frac{1}{|T_2|} \sum_{t \in T_2} a_t^*(\bar{\mu}_1^1) < \frac{1}{|T_2|} \sum_{t \in T_2} a_t^*(\bar{\mu}_1^2) = \bar{a}_2^*(\bar{\mu}_1^2) \tag{32}$$

almost surely. Thus, $\bar{a}_2^*$ is strictly increasing in $\bar{\mu}_1$. It follows that $\Delta\bar{a}$ is a strictly increasing function of $\Delta\bar{\mu}$ almost surely holding $\bar{\mu}_0$ and $\bar{a}_0^*$ constant, so the testing rule in Section 6.2 remains valid under this extension.