The Impact of Input Inaccuracy on Leveraging AI Tools: Evidence from Algorithmic Labor Scheduling

Caleb Kwon*, Antonio Moreno and Ananth Raman

This Version: October 22, 2023

Abstract

Are the inputs used by your AI tool correct and up to date? In this paper, we show that the answer to this question: (i) is frequently a "no" in real business contexts, and (ii) has significant implications on the performance of AI tools. In the context of algorithmic labor scheduling, we propose, identify, and study a problem relating to inaccurate employee availability records, which are used by an AI tool to assign employees to shifts that are necessary to meet required service levels. We study this problem using granular data covering multiple retail chains, which contain more than 74 million shifts that are scheduled for more than 290,000 employees in more than 5,900 brick-and-mortar store locations. In our data, we find that employee availability records are often set incorrectly. Specifically, we find that employees who are no longer available to work are scheduled to work by the AI tool, and employees who are available to work have no existing availabilities. We find evidence that such input inaccuracies directly affect the number of overrides as managers rectify these errors, but also have a spillover effect on shifts that are not subject to input inaccuracies. Ultimately, we find that input inaccuracies take up significant managerial time and have a negative effect on the quality of work schedules, which may lead to a decrease in store performance. Overall, our findings suggest that poor AI input quality management could be one explanation behind the well-documented human distrust of algorithms and the lack of observed business gains in their use.

*[*]**Disclaimer:** This paper makes use of proprietary administrative data on various companies. The conclusions drawn from the data are those of the researchers and do not reflect the views of any companies examined in this paper. These companies are not responsible for, had no role in, and were not involved in analyzing or developing the hypotheses or results reported herein. Caleb Kwon has received financial compensation for consulting work from some suppliers to the companies examined in the paper. He has not consulted for, or received compensation from, the companies examined in this paper.*

1 Introduction

The last decade has witnessed rapid progress in the development of AI tools that help businesses increase operational performance and cut costs (Brynjolfsson and McElheran, 2019). Such tools are used in various contexts including in inventory management (Kesavan and Kushwaha, 2020; van Donselaar et al., 2010), labor scheduling (Kwon et al., 2023), healthcare (Byrne et al., 2023; Dai and Abramoff, 2023; Lin et al., 2022; Liu et al., 2022), employee hiring (Hoffman et al., 2017), product assortment (Kawaguchi, 2021), and in the service industry (Cui et al., 2023; Davenport et al., 2020; Huang and Rust, 2018).

A key question in this literature is the optimal degree of human participation in processes that are now partially or fully automated by the use of AI tools. Most of this debate has focused on *ex post* participation. Specifically, there is some debate as to whether human operators of AI tools should be allowed to participate in the decision-making process *after* having observed the output of the AI tool. This form of ex post participation commonly manifests itself as an override, where the human operator modifies or rejects the recommendations or outputs of the AI tool (Dietvorst and Bharti, 2020; Dietvorst et al., 2015). Participation in the form of making overrides is valuable if human operators have unique insights or private information that the algorithm lacks or cannot adequately process. However, the evidence on the effects of such overrides when examining data from real business contexts is mixed, and appears to depend on the underlying operational context. For example, recent work by Kwon et al. (2023) shows that overrides made by store managers to an AI tool that forecasts demand and schedules labor have positive effects on store performance. In contrast, Kesavan and Kushwaha (2020) study the effects of overrides made to a decision support tool that makes product removal decisions, and find an overall negative impact of overrides. We discuss other papers that have studied ex-post human-algorithm interactions using real-world data in Section 1.1. A key insight from this growing literature is the potential importance of private knowledge (over the algorithm) as one mechanism by which humans can improve upon the decisions made by AI tools.

However, an area that has received relatively less attention is the role of *ex-ante* involvement of human operators of AI tools (We provide a literature review in Section 1.1). One form of ex-ante involvement is the selection and management of inputs into the AI tool or algorithm. These ex-ante interactions may be important in determining the overall performance and effectiveness of AI-driven processes. For instance, the accuracy of demand forecasting algorithms hinges on the precision and comprehensiveness of historical data. If a human operator omits essential data or introduces inaccuracies, it could detrimentally impact the algorithm's predictive capabilities and the subsequent operational decisions (e.g., stock too much or too little inventory). Despite its potential significance, there has been relatively limited empirical or theoretical exploration of ex-ante interactions in the use of AI tools (to our knowledge).

In this paper, we propose, identify, and study the effects of a novel managerial problem centered on ex-ante interactions in the human-AI interface. This problem relates to the possibility that human operators are inputting inaccurate records into the AI tool. We study this problem in the context of algorithmic labor scheduling. In our context, a commercial AI tool automates the demand forecasting and labor scheduling processes across stores.

However, a key task of the human operator (in our case, the store manager) is to set the availability of their employees correctly (i.e., on what days and times each employee can work). Despite how unlikely or hypothetical this issue may seem (indeed, our data providers were unaware of this issue), we show that this step is often overlooked, and that it can: (1) Increase the magnitude of ex-post overrides (identifying both direct and spillover effects), and (2) Decrease the quality of work schedules from both the perspective of corporate officers and employees.

In the context of AI-driven labor scheduling, we refer to this problem as "Employee Record Input Inaccuracy" (ERII). We define ERII as the discrepancy between the actual availability of employees and the records that serve as inputs to the scheduling algorithm. We identify ERII by examining real-world data from multiple retailers that utilize a commercial AI tool to forecast demand and labor scheduling. Our dataset is highly granular, allowing us to identify ERII, measure its prevalence, and examine its impact on the quality of work schedules. Specifically, we observe algorithm inputs pertaining to employee availability, where the algorithm uses these inputs to determine whether an employee is available to work on a given date and time period. We term this data as our "records table". Managers are responsible for maintaining the accuracy of this table. Later when describing the AI tool in Section 2, we show that updating this table is a relatively straightforward process that is not excessively time-consuming. Next, we observe work schedules that are generated by an AI tool, along with the final work schedules given to employees (which capture any overrides made by managers). Finally, we observe human resource records of all employees, which include attributes such as tenure and wage. This table also indicates whether an employee is available to work at a given store. We refer to this table as our "HR table". In total, our data spans more than 74 million shifts, encompassing over 290,000 employees across upwards of 5,900 brick-and-mortar retail locations.

In our context, while the AI tool generates the first batch of schedules, managers have full discretion to modify, add, or manually delete shifts before they are ultimately distributed to employees. We measure ERII by checking every shift that was generated by the AI tool or distributed to employees. In this process, we identify two types of ERII. First, we identify shifts that were generated by the algorithm, but were deleted by the manager because the employee was not actually available to work. This is made possible due to our HR table, which in such instances, shows that the employee has temporarily or permanently left the store. Second, we identify shifts that were not generated by the algorithm but were manually added by the manager due to the employee's availability not having been inputted into the AI tool. The fact that an employee received a shift implies that they are, in principle, available to work. We refer to the former as ERII-Delete, and the latter as ERII-Add. Both arise from a conflict between our "records" and "HR" tables. Additional details about our measurement approach, and specifically why they may be conservative lower bounds, are provided in Section 3.2.

Using our data and measurement approach, we present four key findings. First, despite the potentially hypothetical sounding nature of ERII, we show that ERII is a widespread problem across all retailers and their stores. Specifically, we find in our sample that over 6.1 million shifts (approximately 8.34% of all shifts) are subject to some form of ERII. The wide prevalence of ERII may suggest a systemic gap in managerial practices or perhaps a lack of awareness of the consequences of these inaccuracies on AI-driven processes. We also find this problem to persistent over years, which suggests that ERII is not artifacts of a few anomalous weeks or time periods that result from systematic failures of the AI tool. We refer to these ERII-subject shifts as the direct effects of input inaccuracies. Finally, we find significant cross-sectional heterogeneity in ERII across stores, and that many stores in our sample exhibit very low rates of ERII. This latter fact demonstrates that ERII is not inherent to the quality or design of the AI tool, and that it is entirely possible for stores to be free from ERII.

Second, we find evidence of spillover effects. Specifically, the presence of ERII-subject shifts causes managers to make overrides to shifts that are *not* subject to ERII. Estimating fixed effects models, we find that a 1% increase in ERII-Add and ERII-Delete contributes to an 1.9% and 1.3% increase in overrides to shifts, respectively, that are *not* subject to our definition ERII. Third, we demonstrate that correcting erroneous work schedules caused by ERII consumes a significant amount of managerial time, accounting for both its direct impact and spillover effects. Finally, we find that ERII negatively affects the quality of work schedules. In Section 7, we define two types of schedule quality. The first attempts to capture the quality of work schedule from the perspective of corporate officers. Here, we quantify the misalignment between scheduled labor minutes and the levels that are deemed optimal by the AI tool. We measure this misalignment in granular 15-minute intervals, and we find that this misalignment is larger in the presence of shifts subject to ERII. We also measure the quality of the work schedule from the perspective of workers. To do so, we construct variables that capture the week-to-week consistency of work schedules. For example, we count the number of shifts that are scheduled on a day of the week (e.g., Monday) for an employee that they did not work on in the previous week (Kwon and Raman, 2023a; Lu et al., 2022). The hypothesis here is that humans are less systematic in crafting work schedules, as it is more difficult for humans to maintain consistency and optimize the balance between labor costs and operational demands compared to automated AI tools. Consistent with this hypothesis, we find that ERII has a negative effect on all of our measures of work schedule quality.

In summary, our paper highlights the important role of *ex-ante* human interactions in the human-AI interface. We contribute to the growing literature on the human-AI interface by demonstrating the importance of input quality management on behalf of the humans operating the AI tools. A commonly raised criticism of AI tools is that despite their widespread adoptin, a relatively small percentage of companies see significant financial gains with their use (Ransbotham et al., 2020). Our results suggest that human operators may struggle to fully exploit the benefits of AI tools due to improper management of algorithm inputs. In addition, there is a vast literature that attempts to explain why human operators do not follow the decisions made by AI tools (often referred to as "algorithm aversion"). Our paper suggests that one key factor contributing to algorithm aversion is that input records are not set correctly.

Although our paper focuses on a singular operational context (i.e., algorithmic labor scheduling), our results may be generalizable to other operational contexts where AI tools are used. For example, AI tools are sometimes used to make product selection decisions (Kesavan and Kushwaha, 2020). In such contexts, the accuracy of historical sales data and product attributes is critical. If managers omit crucial product data, or input incorrect ones, it may lead to the selection of less popular products, which may result in decreased sales increased inventory levels. Another example is using AI tools to hire employees (Hoffman et al., 2017). Omitting important data from resumes, or inputting incorrect ones may lead to suboptimal hiring decisions, which may negatively impact team dynamics and overall productivity. We also note that our empirical context focuses on *objective* input errors. Specifically, these errors relate to whether an employee is actually available to work at a given store during a certain period of time. However, there may be other contexts where input decisions are more subjective, such as assessing an employee's suitability for a specific task or role. In such cases, the line between "good" and 'less good" input choices may be less clear-cut, adding another layer of complexity to ex-ante human-AI interactions.

Overall, we demonstrate the existence of a significant problem in the use of AI tools. The key managerial takeaway from our paper is that corporate officers may benefit from establishing clear guidelines and offering training protocols to ensure that their managers accurately input data when using AI tools. In our context, ensuring that the availability records are set correctly can increase alignment with the AI tool's output and increase the quality of work schedules.

1.1 Related Literature

Our paper contributes to the literature on human-algorithm interactions, focusing on the impact of managerial interventions and overrides on operational performance. In this literature, a paper closely related to ours is Kwon et al. (2023), which studies the effects of ex-post overrides made to an algorithm that forecasts demand and schedules labor (identical to ours), and finds that overrides have positive effects on store performance. Another closely related paper is Kesavan and Kushwaha (2020) who use a field experiment to study the effects of overrides in the context of product removal decisions at an automobile replacement parts retailer. They find that overrides reduce profitability by 5.77%, but they also find that the benefits of overrides in specific contexts depend on the type of product being removed.

There are many other empirical papers that study the interactions between human operators and AI tools in real-world business contexts. As discussed earlier, these papers exclusively focus on ex-post interactions (in the form of overrides). For example, van Donselaar et al. (2010) show that manager overrides of SKU replenishment algorithms can increase inventory costs. Second, Ibanez et al. (2017) present evidence that doctor overrides of task-scheduling algorithms can result in decreased productivity. Third, Sun et al. (2022) document that warehouse worker overrides of box-packing algorithms can lead to longer packing times. Fourth, Kawaguchi (2020) find that vending machine manager overrides of assortment algorithms result in reduced revenue. Fifth, Hoffman et al. (2017) find that managers who appear to override test recommendations end up with worse average hires. Finally, Caro and Tejada Cuenca (2018) find that retail store manager overrides of price markdown algorithms can decrease store revenue. We contribute to the literature on human-AI interactions by highlighting a novel form of human-algorithm interaction: the *ex-ante* engagement that involves the manager's involvement in determining the selection and quality of algorithmic inputs.

Our paper is also related to the growing literature that studies the *drivers* of ex-post overrides. Due to the potential negative effects of overrides on operational performance, and the large time opportunity costs of making overrides, understanding the underlying motivations and conditions that lead to such decisions is important for improving upon AI-tool across core operational processes. Using data from a real business context, Caro and Tejada Cuenca (2018) show two common biases that can affect managers in their decision to make overrides in the revenue management context: salience of inventory and sales, and cognitive workload. Other papers have used lab experiments to identify drivers. In a service context, Snyder et al. (2022) finds that participants facing a high customer load rely significantly more on the algorithm's decisions compared to those facing a low customer load. Balakrishnan et al. (2022) find that lab participants take a weighted average between their own prediction and the algorithm's, with a constant weight across prediction instances, regardless of whether they have valuable private information. Finally, DiSorbo and Ferreira (2023) show that warnings on outlier data points – and endorsements on inlier data points – help improve human use of an algorithm that is sometimes fallible. We contribute to this literature by providing a novel driver for ex-post overrides: human error. Specifically, we find that poor input quality management and erroneous inputs can significantly drive the propensity for managers to override AI-generated recommendations. In our context, correcting the input records for a single employee takes very little time (elaborated in Section 2.3). Despite this fact, we still find the frequency and magnitude of input record inaccuracy to be large and persistent.

This paper also contributes to the growing literature on the impact of data quality on the efficacy of AI tools. Here, a closely related paper is Lebovitz et al. (2021), which examines the use of AI in hospitals and finds a disconnect between the tools' theoretical performance and real-world outcomes. Through interviews, the authors pinpoint the crux of the issue as the reliance on ”know-what” knowledge in the training data, advocating for the integration of practical ”know-how” to bolster the tools' practical applicability and accuracy.

Finally, our paper also shares a similar theme to work by DeHoratius and Raman (2008) who show the existence of significant discrepancies between actual and recorded inventory levels. The similarity in theme relates to the mismatch in beliefs and reality of key operational assets, where this mismatch can have negative effects on operational performance. We show that this mismatch in beliefs and reality can also transcend from humans to AI tools. By demonstrating the effects of inaccurate employee availability records on the quality of work schedules, our results support their claim that incorporating up-to-date information is tantamount to achieving operational efficiency.

2 Data and Empirical Context

2.1 Focal Retailers

Our empirical analysis examines data from multiple independent retail companies[1] operating more than 5,900[2] brick-and-mortar store locations in the United States that employ more than 290,000 employees[3]. Our sample begins in 2018 and ends in 2023, but we do not have the same length of data for all retailers. The stores in our sample are present in more than 2,000 cities across all 50 states in the U.S. To maintain the anonymity of the

[1]By independent, we refer to companies that do not belong to same corporate entity. For example, Albertson's, which is not part of our sample, operates multiple retail chains such as Acme Markets and Jewel-Osco. Under our definition, Acme Markets and Jewel-Osco are not independent.

[2]This value represents a lower bound of the number of stores.

[3]This figure includes past and present employees that have worked in the stores in our sample, which we elaborate on in Section 2.4

underlying retailers, we do not disclose the breakdown of the number of stores and employees across retailers. However, we obtain complete and uncensored data for all stores[4] for each retailer. Furthermore, we have detailed information on the labor scheduling processes across stores, including insights into the scheduling algorithm utilized in each store (more details are provided in Section 2.2).

Within their respective companies, stores are largely homogeneous with respect to their operations. This is likely because they are chain stores, with strong brand identity. They record the same number of employee tasks, including facility cleaning, cash register operation, receiving and stocking inventory, customer service, and organizational tasks related to paperwork. However, there are some variations that we observe in terms of products and services that are sold across stores within their respective organizations.

Each store has a single "general manager" that oversees all store operations, making critical decisions related to staffing, scheduling, and overall store management, under the umbrella of the company's broader operational guidelines. As we describe below, the manager has a critical role in the scheduling process across stores as they have the ability to make overrides to the schedules generated by the AI tool. They are also tasked with setting up the availability records of their employees, which are used as inputs by the scheduling algorithm.

2.2 Algorithmic Labor Scheduling

In addition to being homogeneous in operations and organizational structure within their respective firm, all stores in our sample use the *same* commercial[5] AI tool to automate the scheduling of shifts for all employees, including both managerial and non-managerial staff, as well as part-time and full-time staff. Henceforth, we use the term "AI tool" and "algorithms" interchangeable. In our context, a "shift" refers to a specific time interval that outlines an employee's start and end times (hence, duration), break periods, and assigned tasks. We note that while the same commercial algorithm is used across all stores in our sample, there may be some differences in both the underlying parameters and processes across and within retailers. For example, the parameters defining "peak hours" could differ between stores located in high-density city centers compared to those in low-density suburban areas, which may affect how the algorithm schedules shifts to meet customer demand during these distinct periods. Additionally, depending on the size of the store, the optimal mix of labor products between tasks (e.g., cashier versus inventory stocking) may differ between stores within the same retail chain.

The scheduling algorithm has two primary objectives: (1) forecast store demand, and (2) schedule labor to meet a targeted service level based on the demand forecast. To meet the first objective of forecasting demand, the AI tool employs a proprietary machine learning technique that uses a broad range of historical inputs, including past sales and work schedules. The algorithm forecasts demand in 15-minute increments. Based on the predicted demand in the 15 minute interval, the algorithm generates shifts that ensure that the pre-determined service level is met. In addition, the set of scheduled shifts generated by the algorithm simultaneously satisfies employees' labor contracts, employee availability constraints, all relevant labor regulations, and stores' labor budgets. Finally,

[4]Note that the data consists of various unbalanced panel datasets, which are unbalanced solely due to the entrance and exit of stores and not data censoring.

[5]By "commercial", we mean that this algorithm was not constructed in-house, but was provided for payment by a third-party supplier. It is possible that many other retailers and non-retail businesses use this algorithm as well.

we note that the algorithm generates shifts in weekly increments, approximately 2-3 weeks in advance.

We make two additional points about the underlying AI tool. First, the underlying parameters and processes are set and modified by corporate officers, and never by the store manager. Consequently, overrides to work schedules generated by the AI tool also reflect an override to corporate officers, which is a point of concern as it indicates potential misalignment or mistrust between store-level management and corporate directives. This concern was higlighted over a decade ago by Netessine et al. (2010), and then recently examined by Kwon et al. (2023). We elaborate on the potential problems of overrides made to AI tools that forecast demand and schedule labor in Section 5. Second, we note that the algorithm does not "learn" from its own performance (i.e., from demand forecast Mean Squared Errors), or from the overrides that are made by the store manager. More specifically, the algorithm does not condition on its past performance in forecasting demand or scheduling labor, past interactions with specific managers, or expectations about what the store manager will do with its output. In the section below, we discuss how store managers participate in the scheduling process by setting the availability records of their employees, and potentially making overrides to the work schedules that are generated by the algorithm.

2.3 Human-Algorithm Interactions

In our context, the human-algorithm connection is dynamic, with two points in time at which the manager can "interact" with the algorithm. First, the store manager interacts with the algorithm by specifying inputs that are related to the availability of the employee. Here, the inputs are quite granular, with the manager having to specify for each day of the week, and for each employee: (1) when the employee can begin work, and (2) how long the employee can work. As elaborated in Section 2.4 where we describe our data, we observe all the availability records for all employees. Finally, we note that setting the availability records for employees is a relatively straightforward and non-time consuming process. The corresponding author of this paper has observed this process multiple times, and notes that a very conservative upper bound in changing the availability records of a single employee is one minute (provided that the changes are known in advance).

Second, the manager can interact with the algorithm by making edits, or *overriding*, the schedules generated by the algorithm. We refer to the first and second set of interactions as being *ex-ante* and *ex-post*, respectively. While paper's focus is on ex-ante human-algorithm interactions (i.e., specifying the inputs related to employee availability), our a key component of our empirical analyses is how ex-ante interactions can affect ex-post interactions. The latter set of interactions are the primary focus of Kwon et al. (2023), which studies ex-post overrides using data from a single grocery retail chain that uses a similar AI tool.

2.4 Data

Here, we outline the different datasets that we use in our empirical analysis. We describe each dataset and its purpose in our analysis. We conclude with summary statistics.

2.4.1 Algorithm and Manager Work Schedules

The primary dataset is our shift-level dataset containing more than 74 million employee shifts for more than 290,000 employees. In our context, a shift is a time interval of work (usually 4 or 8 hours) that is assigned to a single employee on a given day. Each shift is characterized by five components: (1) the start date, (2) the start time, (3) the duration, (4) the employee, and (5) the assigned tasks (there may be multiple tasks assigned within a single shift). All five characteristics are mutable by the store manager.

We observe two versions of each shift[6]. First, we observe the work schedules that are generated by the algorithm (Version 1, or "V1"). Second, we observe the work schedules that are ultimately distributed to the employees (Version 2, or "V2"). The differences between the work schedules between V1 and V2 are fully determined by the ex-post overrides that are made by the store manager. These overrides can take the form of edits to any of the five characteristics listed previously[7]. Overrides can also take the form of shifts being manually deleted or added between Versions 1 and 2. We therefore define an override as: (1) an edit to any of the five shift characteristics, (2) a shift being manually deleted between V1 and V2, or (3) any of the five characteristics being changed. This definition is consistent with how the focal retailers define overrides.

2.5 Employee Availability Records

The second dataset contains the precise records of employee availability that are ultimately used as inputs by the AI tool. This dataset contains an exhaustive record of an employee's availability at a particular store at any given time. Here, the unit of observation is at the worker-store-period level. Multiple observations at the person-period level can exist if availabilities are updated. As we have stated above in Section 2.3, the updating process is straightforward and non time-consuming.

Each observation contains availability information for all seven days of the week. Among other variables, we observe at the worker-store-period-day of the week level: (1) the earliest time at which an employee can commence their work (for example, 9 AM), (2) the maximum duration that the employee is available to work (for example, 8 hours), and (3) any restrictions on the timing or length of the break periods during their shift.

2.6 Human Resource Records

Finally, we have access to the human resource records for each employee in our sample. These records, maintained by corporate-level human resource officers for payroll and legal compliance, are not under the active management of store managers. Within this dataset, data is structured at the employee-store-period level, capturing details such as: (1) hire date, (2) termination date (if applicable), (3) wage, (4) gender, (5) job title, and (6) employment status (part-time or full-time). An employee can have multiple entries for reasons including transfers between stores, promotions, changes in employment status, or wage adjustments. As elaborated in Section 3.1,

[6] In practice, we also observe a third version, which is the version that is sent to payroll. But conditional on the second version existing, the third version is almost always identical (over 99%).

[7] We note that because the employee is a characteristic of the shift, the final work schedule that is distributed to employees can be accomplished in different ways. For example, a manager can move an employee from a shift that was algorithmically generated to start at 9AM to 10AM by simply changing the start time. However, they can also achieve the same result by removing the shift at 9 AM and creating a new one at 10 AM (with the same values for the other four characteristics).

this dataset is critical for identifying inaccurate inputs because it offers a "ground truth" concerning the availability of employees. Specifically, it reveals the effective start and end dates of an employee's tenure at a particular store, indicating that they are actually available to work. Availability is based on their employment contract, and the fact that payroll is active for that employee. We refer to an employee being available in this records table as being "active", which is a necessary condition for being able to work. Sufficiency is not guaranteed, for instance, because an employee could communicate to their manager that they are ill or unavailable.

2.7 Summary Statistics

We present summary statistics of all relevant variables in our empirical analyses at the shift and store levels. Table 1 presents these summary statistics. We make a few observations here, and review these statistics in greater detail when examining the magnitudes of ERII in Section 4. Panel A presents our summary statistics at the shift-level ($n \geq 74$ million). Here, we note that the rates of overrides are large. When combining instances of shift additions, deletions, and modifications of the shift (of any of the five characteristics for each shift), the rate of shift-level overrides is 83.7%. These statistics echo the findings of Kwon et al. (2023), who show that the frequency and magnitude of overrides are large for a grocery retailer that uses a similar AI tool. Understanding and addressing the causes and effects of ex-post overrides is a priority for retailers. Consequently, one of our objectives in studying ERII is to determine whether they can increase the magnitude of overrides, and whether they have a material effect on the quality of work schedules (which have plausible connections to store performance).

Next, Panel B presents our summary statistics at the store-date level ($n \geq 7.7$ million). Here, we note that as a result of overrides between V1 and V2, the alignment between forecasted demand and scheduled labor decreases by 19.5 percentage points. As we describe below, this alignment is a key metric that corporate officers examine when assessing the quality of overrides that are made by managers in their stores. Finally, Panel C presents store-level (cross-sectional) summary statistics. The fact that this alignment is decreasing, on average, between V1 and V2 is another reason why corporate officers are concerned about overrides. That is, because corporate officers are involved in the demand forecasting process[8], overrides to V1 schedules made by store managers, are in a sense, overrides to the decisions made by corporate officers. We use these summary statistics later in Section 4 to demonstrate the large cross-sectional variation in the frequency and magnitude of ERII.

3 Inaccurate Input Records: An Impediment to Leveraging AI Tools

In this section, we propose and define a problem related to the quality of input records that may be a significant impediment when leveraging AI tools to automate or improve upon core operational processes. As discussed earlier, a commonly raised concern about AI tools is that despite their widespread adoption, a relatively small percentage of companies see significant financial gains (Ransbotham et al., 2020). In response to this finding, we raise the possibility in this paper that these modest returns may not be a reflection of the intrinsic capabilities of AI tools, but rather, they may stem from human-introduced inaccuracies in the data input process.

[8] As explained previously, they are involved in the selection of inputs and also decide on the values of key parameters.

Despite its colloquial nature, the adage "garbage in, garbage out" is a suitable description of the problem with inaccurate inputs when utilizing AI tools: Even the most advanced AI systems can perform poorly if the input data on which they operate are inaccurate. For instance, consider AI-powered demand forecasting tools. The effectiveness of these systems hinges upon the volume and accuracy of past sales data to predict future demand. However, if inputted historical sales records are erroneous, or if an insufficient amount of data is inputted, the AI's forecasts may lead perform poorly. This poor performance may have negative operational consequences in the form of overstocking due to an incorrect high demand forecast or stockouts due to an incorrect low demand forecast. Inaccurate input records can occur in this context, for example, if the operator of the AI tool mixes up sales from one department with another (e.g., shoes versus handbags). Similarly, AI-driven SKU (Stock Keeping Unit) selection tools, which help retailers optimize their product range based on metrics such as historical sales and market trends, can perform poorly if provided with inaccurate customer preference data. This could result in retailers holding onto unwanted stock or not obtaining products that have high demand during key holiday sales. Finally, in the domain of employee hiring, AI tools are often used to select candidates and predict the success of the job role. However, if resume data are incorrectly inputted into the algorithm, this oversight may cause the AI tool to overlook qualified candidates and select unqualified candidates, which may result in high turnover and increased training costs. Collectively, these examples highlight the critical importance of input record accuracy, and may be a significant factor in why firms are not seeing the expected financial returns from AI tools.

Given the simplicity of these errors, the gravity of their potential effects, and the straightforwardness of their remedies, the examples above may seem strongly hypothetical in nature. Managers are frequently evaluated based on the performance of their stores, and thus there are clear incentives to effectively utilize the AI tools provided by corporate officers. Furthermore, corporate officers commonly express concerns associated with the magnitude of overrides, and input errors may lead to the generation of erroneous work schedules, which must be rectified by making overrides. However, we show in our analyses below that a similar (and simple) problem of input inaccuracies is prevalent and persistent within another core operational process: labor scheduling.

It is well known that labor scheduling is commonly automated by AI tools among large retailers (Kwon et al., 2023; Netessine et al., 2010). The problem of inaccurate inputs in the context of schedule labor relates to the true availability of workers, which may conflict with the inputs that are fed into the AI tool. For example, due to inaccurate records, an employee who no longer works for the company may be erroneously generated a work schedule by the AI tool, and an employee who can work may be scheduled without work. In addition, due to the interconnected nature of work schedules (i.e., the combination of work schedules across employees sets a pre-defined service level), such inaccuracies can cascade and spillover onto the work schedules of employees that are not subject to such input record errors, leading to inefficient staffing, operational disruptions, and unmet service demands.

In this paper, we show that the problem of input inaccuracies when using AI tools to generate work schedules in retail stores is: (1) large, (2) persistent, and (3) have large negative effects on the quality of work schedules. In the remainder of this section, we provide our definition and approach for measuring inaccurate input records in the context of scheduling labor using AI tools.

3.1 Definition of ERII

We define ERII as the discrepancy between the actual availability of employees and the records that serve as inputs to the scheduling algorithm. The basic intuition is as follows: Just because an employee is generated a V1 work schedule by the AI tool does not mean that they are an active employee of the store. Similarly, an employee who is active and who receives a V2 work schedule may not have an employee availability record. Such discrepancies can result in overscheduling or underscheduling of employees. These discrepancies are discovered through a digital inspection of two data tables that are managed independently: (1) Employee Availability Records and (2) Human Resource Records. These tables were introduced and discussed in Section 2. Below, we discuss how these tables can be used to identify ERII.

First, for every shift in V1 and V2, we specify its corresponding match to the data table associated with employee availability records that are fed into the algorithm as inputs. The fact that a shift was generated by the AI tool means, in principle, that there exists *some* availability record. We recall that this particular data table is maintained at the store level (by management). These records contain observations at the worker $\times$ store $\times$ time-period level. Each observation indicates at what periods of the day, and for how long, a given employee at a particular store can work. Each observation has an effective and end date indicating the duration of time that the record is valid. For example, between January 1st, 2023 and June 1st, 2023, John Doe at Store 123 can work from 9AM to 5PM from Mondays to Fridays, with no availability during the weekends. This table contains additional details, such as the minimum and maximum number of hours that the employee can be scheduled per week.

Second, to verify the status of each employee, we inspect the data table that covers human resource records, which are maintained at the corporate level. These records are not as granular as the employee availability table. At the employee $\times$ store level, they only indicate the employee's effective and end date at the store. That is, they do not provide information on which days of the week or how long the employee can work. We are told that this table is accurate because due to corporate policy, managers are required to inform HR when an employee leaves the store or company. In addition, they need to be active in this table to be paid for their work. Finally, HR is directly involved in all employee hiring, which further ensures the accuracy of these tables. No such mandates are made regarding the availability records to the AI tool.

For the aforementioned reasons, we take the HR data table as the "ground truth" when measuring ERII. It is a necessary and sufficient condition for an employee with an active status in a store (i.e., available to work) to be active according to HR records. In contrast, having an active employee availability record is neither sufficient nor necessary for an employee to be available to work at a given store. For instance, HR can onboard an employee, but the store manager may simply forget to add that employee's availability records into the algorithm. Consequently, erroneous work schedules that are generated by the algorithm are fully caused by inaccurate employee records, and *not* HR records.

As depicted in Figure 1, an inconsistency between the employee availability records and the HR data can lead to various scenarios. When both data tables are consistent, with respect to an employee's status at a given store, the employee is successfully scheduled without issues (top left box). In cases where an employee is not present in the HR data but is in the availability records and is available to work, the manager must manually add the employee to the set of work schedules. Conversely, if an employee appears in the HR data but not in the availability records, they are manually removed. Lastly, if an employee is not found in either table, this indicates that the employee no longer works at the store, and is (correctly) not generated or provide a work schedule.

3.2 Measurement of ERII

While the definition of ERII is straightforward (i.e. inaccurate input records with respect to employee availability), its measurement is more nuanced in our context. This is because not all erroneous input records lead to infeasible schedules. To be concrete, suppose that an employee, John, quits. But, by accident or forgetfulness on the part of the store manager, his new availability (or lack thereof) is not correctly updated in the availability records that are fed as inputs into the algorithm. However, despite the inaccurate input of John's availability, there may still be no instances of erroneous work schedules because the scheduling algorithm, due to a low demand forecast, did not schedule any work for John. This scenario can occur if the store has a large surplus of staff or if the scheduling algorithm prioritizes other employees over John due to their skills, seniority, or other constraints.

Therefore, incorrect input records lead to a *distribution* of incorrect work schedules, where this distribution is influenced by other variables such as the demand forecast (how much labor is needed), the total labor supply, the type of tasks needed, and other demand and labor related variables. Therefore, the actual realization of an infeasible work schedule due to ERII is not solely dependent on the inaccuracy of the input, but also on how this inaccuracy interacts with the broader work environment.

Consequently, the set of *realized* erroneous schedules is a probabilistic mapping to ERII. We measure ERII in two ways:

1. **ERII-Delete**: In the set of V1 shifts that are subsequently deleted (i.e. no corresponding V2 shift exists), we check whether each shift is associated with an employee that has an "active" status at the store according to our HR table. If no such record exists for that employee, this means that, in principle, the employee is no longer able to work at that particular store, and that the employee's shift was deleted because of inaccurate availability records. We designate such shifts as being subject to ERII-Delete.

2. **ERII-Add**: In the set of V2 shifts, we check whether each shift is associated with an employee who has *any* availability in the input records table. If no such record exists for that employee, this means that the employee's availability records were not correctly inputted. The fact that the employee received a V2 shift implies that, in principle, they are able to work. However, the fact that there are no availability records for the employee implies that the algorithm could never have included that employee in its set of V1 shifts. We designate such shifts as being subject to ERII-Add.

We prefer to use realized instances of ERII (described above) over simply using the number of inaccurate input records in our analyses for two key reasons. First, inaccurate records map to a distribution of erroneous schedules (e.g., based on the demand forecast), but it is unclear how to characterize this distribution. For example, there is insufficient information in the HR table to predict how many shifts an employee should have been assigned, had their availability records had been correctly specified. For instance, the magnitude of ERII will depend on the forecasts made on a given day, but the mapping between the demand forecasts and the unknown is known to us. Accordingly, it is only through the managers actions of manually adding employees with an active HR record (but without an input record) that we can assess the magnitude of this problem. Second, realized instances of ERII may more closely capture the true effect on the quality of work schedules and other operational metrics, whereas a mere count of inaccurate records may inflate the perceived problem of ERII without having real effects on work schedules or on the time opportunity cost of managers.

3.3 Our Measures of ERII are Conservative

Before moving onto our empirical analyses, we make the final point that our measures of ERII are conservative for several reasons. First, we note that our ERII measurements capture "large" input record inaccuracies where the employee is incorrectly specified to have *zero* availability (ERII-Add), or has some availability when they actually have none (ERII-Delete). These are two polar cases that do not capture intermediate forms of input record inaccuracy (described below) that may also be frequent in the data. Second, we note that our ERII measures only capture input errors associated with *availability*, and not based on assigned tasks. Consequently, our measures of ERII do not capture, for instance, instances where an employee was promoted and should now be assigned to a different task.

For clarity about both examples, consider first an active employee named Sarah. By active, Sarah has been fully onboarded, has payroll set up, and is able to work. Sarah's availability records indicate that she can work Mondays through Friday for eight hours each day. However, in reality, Sarah has recently started attending evening classes, rendering her unavailable for late shifts on Wednesdays and Fridays. Despite this change in her schedule, it is possible that due to oversight, forgetfulness, or being preoccupied with other managerial tasks, the manager at her store does not make the appropriate adjustments to her input records (to the algorithm). As a result, she could still be scheduled for the evenings on Wednesdays and Fridays. Accordingly, with some non-zero probability, her work schedules will be subject to ERII because the AI tool may generate work schedules with her working on a Wednesday or Friday evening.

Another example is related to task specification. Consider another employee James, who is highly skilled in inventory management but has recently been moved to customer service due to a perceived shortage there. However, due to managerial oversight, the records still indicate his primary skill as inventory management, and as a result, whenever there is an inventory-related task, the system continues to schedule him on inventory-related duties. This example is different from the case of Sarah because, while James' availability is correctly specified, his task specification is outdated in the system. Therefore, although James is physically available for work, he is scheduled for tasks that do not align with his current role. This discrepancy can lead to a situation where James' skills are not utilized optimally, or where he might be scheduled for inventory tasks while actually needed elsewhere, such as customer service.

4 Empirical Facts on ERII

This section begins our empirical analysis on ERII. Here, we present a series of stylized facts that characterize and quantify ERII in our data. Our primary objectives are to show that ERII: (1) occur with considerable frequency and magnitude, (2) exhibit considerable heterogeneity across stores, and (3) are a persistent problem within our sample that includes all retailers and their stores.

Stylized Fact 1. *Both types of ERII occur with considerable magnitude.*

Our first stylized fact serves to counter the notion that input inaccuracies are merely hypothetical in nature. As stated earlier, the hypothetical nature of ERII is understandable: the retailers underlying our study are large and sophisticated based on its utilization of AI tools to automate various core tasks (beyond forecasting demand and scheduling labor). Additionally, it is reasonable to believe that store managers, who are evaluated based on their respective store's performance, would try to utilize the AI tool as effectively as possible.

However, contrary to this claim, we find that ERII-Add and ERII-Delete occur with significant magnitude and frequency. We present shift-level summary statistics about ERII-Add and ERII-Delete (among other variables) in Panel A of Table 1. Here, we find that in our shift-level sample containing more than 74 million shifts that exist in either V1 or V2 (across all retailers), 4.97% ($\approx$ 3.6 million) and 3.34% ($\approx$ 2.5 million) of shifts are subject to ERII-Add and ERII-Delete, respectively. Combined, more than 6.1 million shifts (8.34%) required manual adjustments (either by adding or deleting) due to inaccurate employee availability records.

To better understand these magnitudes, we aggregate our shift-level data to the store-date level ($n \approx 7.2$ million). We compute four measures at the store-date level: (1) Total ERII-Add Shifts / Total V2 Shifts, (2) Total ERII- Add Duration / Total V2 Duration, (3) Total ERII-Delete Shifts / Total V1 Shifts, and (4) Total ERII - Delete Duration / Total V1 Duration. Panel B of Table 1 presents summary statistics of these four measures across and within our four companies. We find that the mean (standard deviation in parentheses) of these four measures when pooling all stores across our retailers are 0.064 (0.189), 0.071 (0.171), 0.058 (0.180), and 0.068 (0.175). Overall, our first stylized fact demonstrates that ERII ocurrs with significant magnitudes in our sample.

Stylized Fact 2. *ERII is a persistent problem.*

Next, we show that ERII is persistent over time, and that the large magnitudes that we find in the data are not artifacts of a few anomalous weeks or time periods that may result from, say, systematic failures of the scheduling algorithm that affect many stores. Additionally, this stylized fact rules out the possibility that our documented magnitudes of ERII are driven by transient issues associated with the initial implementation of the scheduling algorithm or the acclimatization period needed for managers to familiarize themselves with the system. Indeed, a plausible hypothesis is that managers might struggle to adapt to the new scheduling system, and hence, the initial period might be marked by high levels of ERII as managers understand how the algorithm works and the importance of accurately feeding in correct employee availability records.

To illustrate this point, we present summary statistics of ERII after aggregating our data to the store-year level. From 2018 to 2023, the mean (standard deviation in parentheses) of the percentage of ERII-Add subject shifts is

0.038 (0.026) in 2018, 0.051 (0.085) in 2019, 0.066 (0.090) in 2020, 0.067 (0.114) in 2021, 0.049 (0.107) 2022, and 0.025 (0.093) in 2023. In the same time period the mean (standard deviation in parentheses) of ERII-Delete subject shifts is 0.024 (0.020) in 2018, 0.070 (0.082) in 2019, 0.087 (0.115) in 2020, 0.093 (0.139) in 2021, 0.090 (0.255) in 2022, and 0.063 (0.283) in 2023. We note that we have a much smaller sample in 2018 (less than 10% of the sample size in 2019 – 2022 (inclusive), and our data in 2023 are incomplete, since they end in March. Overall, we find in our data that both ERII-Add and ERII-Delete are persistent issues that have not disappeared over time.

Stylized Fact 3. *There is significant cross-sectional heterogeneity in ERII across companies and stores.*

Third, there is considerable cross-sectional variation in ERII across companies and stores. To demonstrate this fact, we point to our cross sectional statistcs in Panel C of Table 1. We also present a histogram of our four ERII-measures in Figure 2.

First, we note that cross-sectional heterogeneity is evident from the comparison of the mean and standard deviations in the cross section of our stores (Panel C). Specifically, we find that in our sample of more than 5,900 stores, the standard deviation of our four ERII metrics (presented above) are at least as large as the mean values of the measures.

Second, we note from our histograms in Figure 2 that all four measures are significantly right skewed, with considerable mass in the histogram found closer to 0. In fact, in the cross section of our stores, we find that 64.8% of stores have ERII-Add Shifts (as a percentage of V2 shifts) of less than 5%. Additionally, approximately 34.2% of stores are subject to ERII-Add Shifts (as a percentage of V2 shifts) of less than 1%. Finally, about 6.2% of stores are completely free of ERII based on this latter measure. Overall, these statistics demonstrate that it is possible to have very low amounts (or be completely free) of ERII (based on our definitions). That is, the fact that many stores have little to no shifts subject to ERII suggests that this problem may not be inherent to the quality of the AI tool, or the difficulties in updating employee records.

5 Hypotheses on the Real Effects of ERII

In the previous section, we proposed, described, and defined a potential problem related to the quality of input records when using AI tools. In the context of scheduling labor, we referred to this problem as ERII. However, beyond the added time managers spend correcting ERII-related issues, a key managerial question is whether the presence of ERII affects the quality of work schedules and store performance.

In this section, we develop and discuss three hypotheses on the potential effects of ERII on the quality of work schedules and store performance. To illustrate our hypotheses, we draw an analogy between the development of work schedules and a knitted article of clothing (e.g., a sweater), and then use this analogy to outline our hypotheses. The purpose of this analogy is to show that work schedules, much like a knitted sweater, are an integrated product of various interwoven components (threads in the sweater, and shifts in the work schedule); any flaw or irregularity in one part can compromise the integrity and function of the entire piece.

5.1 A Conceptual Understanding of Work Schedule Complexity

At a casual glance, scheduled shifts on a given day can appear mechanistic and isolated. It is easy to perceive shifts merely as a matrix of names, dates, and times, isolated from their potential interdependence in accomplishing key tasks. Even to the assigned employees, how their shifts were developed, along with their overarching purpose in store operations, may never cross their minds due to the more immediate and practical concerns of the work schedules' effects on the employees' personal routines around their assigned shifts.

In reality, the development of work shifts is complex, and its complexity is akin to the interweaving of threads in a knitted piece of clothing, such as a sweater. Based on the blueprint (often referred to as "tech packs" in the fashion industry), the threads in the sweater are knit together in a precise sequence and pattern. In the same way, work shifts are also constructed with a specific blueprint in mind, interwoven with other shifts, and tailored to the objectives of store operations. Each shift serves a distinct purpose (e.g., stocking shelves or working in customer service) to ensure that tasks are completed efficiently and service levels are maintained throughout the work day. Crucially, shifts often exhibit interdependence, as a shift can continue, complement, or start tasks that are assigned to previous, concurrent, or future shifts, respectively. In our context, these work schedule blueprints are the result of a collaborative effort between the corporate officers of each retailer and the developers of the AI tool. The blueprints are customized to each individual store, considering an extensive list of factors such as the layout, size, demographics, and neighboring traffic patterns. The corresponding author of this paper has examined these blueprints, and found them to condition on a wide array of constraints, objectives, forecasts, and historical data.

This analogy lays the foundation for our three hypotheses: it is unlikely that managers can make ex-post overrides to fully rectify work schedules that are developed based on incorrect or incomplete blueprints. Drawing on our analogy to a knitted sweater, it is unlikely that the knitter can

ERII-Add can be conceptualized as discovering additional threads after the sweater is already in the process of being knitted. Initially, these threads were not known to be available, so the sweater was started with a perceived limited set of threads (akin to the algorithm not being aware of the availability of certain employees). Now, with the newfound threads, the sweater may need to be altered or augmented in ways that were not originally planned. ERII-Delete, in contrast, is akin to discovering that some threads, which were believed to be part of the knitting process, are actually unsuitable or flawed. As a result, those threads must be removed, even though they have already been integrated into the sweater. This mirrors the algorithm incorrectly assuming the availability of certain employees and scheduling them, only to realize later that they are not actually available.

Overall, while managers may attempt to rectify issues related to ERII-Add and ERII-Delete through ex-post edits, it is akin to patching or modifying a sweater after it has already been knitted; the result, although functional, may not possess the integrity or efficiency of a sweater developed with a correct blueprint, with full knowledge of available material. With this underlying intuition, we now present and discuss our three hypotheses on the effects of ERII on work schedule quality and store performance.

Hypothesis 1. *ERII Increases Ex-Post Overrides*

Our first hypothesis is that ERII can increase overrides made to shifts that are *not* subject to ERII. These spillover effects can occur for several reasons. First, a shift that is subject to ERII-Delete must be deleted (because the focal employee is not available to work), and potentially replaced with another shift. This latter action can be achieved by moving a shift not subject to ERII to the deleted shift's position, or by generating a completely new one. Both count as overrides in our context. Second, a shift that is subject to ERII-Add can be used to replace a pre-existing shift, which would require the pre-existing shift to either be deleted or moved elsewhere. This can occur, for example, if the ERII-subject employee is a better fit for a particular task.

Another reason that ERII can increase overrides is that they may inadvertently amplify managerial mistrust in the scheduling algorithm, leading to a higher propensity to make overrides. Interestingly, this mistrust can be triggered by the ERII itself, a component within the manager's control.

More than a decade ago, Netessine et al. (2010) highlighted the potential tension between store managers and the work schedules produced by commercial workforce algorithms, which are predominantly calibrated to the preferences of corporate officers. Their citation of a report from Retail Systems Research underscores this tension, noting that "*...store managers did not believe that a corporate team understood what went on in stores well enough to define either labor standards or the basis for calculating labor budgets...*" Furthermore, it was observed that "*...store managers spent very little time on the schedule prior to the implementation... but the impact on employees was terrible*." This historical provides important context for understanding the delicate balance of trust between store managers and scheduling algorithms.

Therefore, due to ERII, a systemic cycle of "bad input management" could exacerbate what is commonly referred to as "algorithm aversion" (Balakrishnan et al., 2022; Dietvorst and Bharti, 2020; Dietvorst et al., 2015; Snyder et al., 2022), where managers may develop a lack of trust in the algorithm due to the perceived need for continuous corrections. However, perhaps ironically, our first hypothesis raises the possibility that some portion of the observed "algorithm aversion" may actually stem from human errors that affect the quality of inputs. In other words, the mistrust and skepticism that managers exhibit towards the algorithm might be self-inflicted to some extent, since it is their inaccuracies in data input that can lead to suboptimal scheduling decisions made by the AI tool. This irony highlights the need for improved awareness and training about the importance of accurate data input, as these human-induced errors could inadvertently lead to the decreased effectiveness, or even abandonment, of potentially beneficial AI tools. Overall, our first hypothesis raise the possibility that some portion of the observed "algorithm aversion" arise based on human errors that affect the quality of inputs.

Hypothesis 2. *ERII has a non-linear effect on Ex-Post Overrides, such that the effect amplifies as the magnitude of ERII increases.*

Our second hypothesis relates to the possibility non-linear effects that ERII may have on overrides. Specifically, we note that overrides that managers make in the presence of ERII do not necessarily have to be one-to-one with the ERII-Add or ERII-Delete shift. That is, the adjustments that are necessary to rectify shifts subject to ERII may have a *cascading effect*, where multiple adjustments are required for a single shift that is subject to ERII. To be concrete, consider an ERII-Add scenario where an available employee is integrated into the set of morning shifts displacing another employee to the afternoon. This displacement can occur, for example, if the

ERII-Add employee can only work during the mornings, whereas the displaced employee can work both morning and afternoon. However, suppose that this adjustment causes a surplus of labor for a particular task, causing another employee that is scheduled at the same time to be moved to another task. Accordingly, even though the trigger was a singular ERII-Add, these spillover effects can cascade through multiple shifts, leading to a series of interconnected overrides.

Hypothesis 3. *ERII Decreases Work Schedule Quality*

Our third hypothesis relates to the real effects on work schedules that result from ex-post overrides that are made due to ERII (both direct and spillover effects). This hypothesis stems from the uncertainty about whether managers, with their ex-post overrides, can consistently produce work schedules that achieve the same level of optimality as those the algorithm would produce in the absence of ERII. Going back to our knitted sweater analogy, it is uncertain whether the various thread adjustments can restore the sweater to the quality of one crafted from an error-free blueprint.

This uncertainty is based on the sophisticated and tailored (to the particular store) nature of the algorithm. Specifically, the algorithm employs machine learning techniques to match forecasted demand with appropriate service levels, considering variables such as historical data, employee availability, and skill requirements. These techniques are tailored to each particular store, and these techniques are formulated through a lengthy process involving multiple entities (corporate officers, the developers of the algorithm, engineers, and consultants). This automated approach allows the algorithm to optimize the scheduling process while considering a wide array of objectives (e.g., service levels) and constraints (e.g., budgets and labor laws). Despite corrective actions by managers, it is unclear whether work schedules subject to ERII match the quality of work schedules that are completely free of ERII.

In our analyses, we use two proxy for the quality of work schedules. From the perspective of corporate officers, we measure schedule quality based on how closely work schedules match the required work levels that are determined by the algorithm. We aggregate deviations between scheduled and required work-levels at 15-minute intervals to capture the cumulative discrepancies over a day. This cumulative metric offers insights into the overall consistency and reliability of work schedules to meet algorithmically determined demands. Second, from the perspective of employees, we proxy schedule quality based on how stable work schedules are from week to week. Both proxies are explicitly defined in Section 7.

6 The Spillover Effects of ERII on Ex-Post Overrides

One of the hypotheses made in Section 5 is that the presence of shifts subject to ERII may have spillover effects on the shifts *not* subject to ERII. To be concrete, consider a situation where a shift has to be added manually for an employee due to ERII-Add at a particular time and for a specific task (a direct effect of ERII). This change, although seemingly isolated, could inadvertently unsettle the scheduling equilibrium by increasing service levels beyond an optimal range. Consequently, to accommodate the addition of this new shift, the manager may need to reassign other employees, originally scheduled at the same time, to different tasks or roles. Alternatively, these employees may be asked to start their shift later or earlier, or they may even experience shortened or extended shifts to balance service-level requirements. These are the *spillover* effects that are described in Hypothesis 1.

Overall, as hypothesized in Section 5, the combined effects of ERII (direct and spillover) on overrides are potentially problematic because: (1) they represent significant time opportunity costs, and (2) it is unclear whether managers can produce work schedules as optimal as those generated by the algorithm.

Finally, we note that examining spillover effects (on shifts not subject to ERII) is critical to avoid circularity because a shift that is added or deleted due to ERII counts, in principle, as an ex-post override. Consequently, in our analyses below, we study the effects of ERII on the subsample of shifts that were *not* affected by our definitions of ERII.

6.1 Empirical Model

To test for the possibility of ERII spillover to shifts that are *not* subject to ERII, we estimate variants of the following linear probability model:

$$Same_{i,t,s} = \beta_1 \cdot \log(1 + ERII\_Add_{i,t}) + \beta_2 \cdot \log(1 + ERII\_Delete_{i,t}) + \gamma \cdot \boldsymbol{X}_{i,t} + \alpha_i + \lambda_t + \varepsilon_{i,t} \tag{1}$$

where $Same_{i,t,s}$ is a dummy variable that indicates where shift s in store i on date t is identical between Versions 1 and 2, i.e., experienced no override. We estimate Equation 1 after restricting our sample to the set of shifts that are not subject to either type of ERII, which are captured by the variables $\log(1 + ERII\_Add_{i,t})$ and $\log(1 + ERII\_Delete_{i,t})$. These variables are logged after adding one to study the effects of a percentage change in the magnitudes of both variables, and where one is added to accommodate for the fact that there may not be ERII at a store on a given date. Next, α_i is a store fixed effect, and λ_t is a time fixed effect. In addition to these fixed effects, we include store-date level control variables, $\boldsymbol{X}_{i,t}$, which include the demand forecast, operating time, and the total amount of labor that was scheduled in V1. Finally, we double cluster standard errors at the store and year-week levels to capture arbitrary correlations at both these levels[9].

Our coefficients of interest are β_1 and β_2 which represent the correlations between the probability that shifts are not overridden (between V1 and V2) and ERII-Add and ERII-Delete, respectively. Based on the included fixed effects, our approach here estimates these associations by examining within store and year-week groups of observations. These fixed effects are important for the following reasons. First, store fixed effects (α_i) account for time-invariant (or slow moving), store-specific factors that could correlate with ERII-Add and ERII-Delete. For example, stores (and companies) may differ in their management style or operational practices, which may be factors that also influence the propensity for manual adjustments to schedules. If one store has a higher tendency to manually adjust schedules due to, say, a store manager's meticulous style, it may also exhibit lower levels of ERII-Add or ERII-Delete due to their careful attention to the algorithm's inputs. Additionally the magnitude

[9]Double clustering at both the store and year-week levels corrects for potential correlations within groups defined by the store, as well as correlations within groups defined by the year-week period. For example, there may be store-specific factors, such as managerial style or local labor market characteristics that may affect hiring and turnover, that cause errors to be correlated across different employees within the same store. Similarly, certain events or macroeconomic trends may cause errors to be correlated across all stores in a given year-week period.

of ERII and overrides in general may also be correlated with store size, which is accounted for by store fixed effects. For instance, it may be more difficult to keep track of employees' availability records at larger stores, and in addition, the AI tool may find it more challenging to forecast demand. Consequently, larger stores may have systematically more overrides and magnitudes of ERII.

Next, year-week fixed effects (λ_t) help to control for time-specific factors, such as macroeconomic variables that may affect overrides and ERII. An example is the seasonality associated with the overall labor market. For example, periods of high labor turnover or increased hiring could be correlated with both the incidence of ERII and the level of manual overrides. During such periods, the need to update employee availability records increases, but managers may overlook or forget these updates, contributing to the prevalence of ERII. Moreover, managers may possess private information about newly hired employees' capabilities or other attributes, which could prompt them to make overrides that the AI algorithm—lacking this nuanced, non-quantifiable information—would not account for.

In addition, to our preferred specification (above), we also consider a specification where we interact store and year-week fixed effects. We do so to capture unobservables associated with stores that are dynamic within stores over time. For example, we mentioned earlier that "better" managers may influence both ERII and overrides. However, store and year $\times$ week fixed effects are insufficient to address this source of bias. In addition, changes to the AI tool (e.g., a change in the service level requirement), an introduction of new protocols, or company-wide training programs aimed at reducing input errors, are not addressed by our additive fixed effects. One downside to this approach, however, is that because work schedules are generated in week increments across all stores (which implies that ERII is also a weekly phenomenon), these fixed effects may absorb some of the variation in overrides that we are attempting to attribute to ERII.

6.2 Results

Table 2 presents the estimates associated with Equation 1. This sample includes more than 69.4 million store-date-shift level observations. The dependent variable across all columns is a binary variable that equals 1 if the shift is *identical* across all five shift characteristics. Therefore, in addition to the manual addition or deletion of a V1 shift, an edit to any shift characteristic (start time, duration, employee, break time, or task) counts as an override. Columns (1) through (5) of this table differ only based on the included fixed effects and control variables. In our preferred specification that includes all control variables and fixed effects (Column (5)), we find that a 1% increase in ERII-Add and ERII-Delete, at the store-date level, corresponds to a 0.9 and 1.9 percentage point decrease in the probability that the shift was identical between V1 and V2 (both estimates are significant at the 1% level). Relative to the sample mean of 0.481, these estimates reflect a 1.87% and 3.40% increase, respectively. Accordingly, these estimates imply that overrides are *more* likely to happen to shifts not subject to ERII-Add and ERII-Delete that are scheduled alongside shifts that are subject to ERII-Add and ERII-Delete, experience higher rates of overrides. We therefore president evidence that ERII causes spillover effects, consistent with Hypothesis 1.

As described earlier, one concern with our current model specification in Equation 1 is that it treats store, and year-week fixed effects as separate and additive, which potentially fails to important interaction effects. For example, a particular store might have a higher degree of schedule complexity due to factors like larger staff size or varying shift demands, which could lead to a higher occurrence of ERII-Add and ERII-Delete events, irrespective of the general propensity of the store's manager to make errors in inputting employee availability. Additionally, specific periods like the holiday season could increase scheduling complexity, exacerbating input errors at an employee-store-week level due to the higher workload and leading to a spike in ERII-Add and ERII-Delete events. Furthermore, endogeneity may arise if certain unobserved factors influencing the likelihood of overrides also correlate with the occurrences of ERII-Add and ERII-Delete. For instance, during the holiday season at a particularly complex store, a manager might be more likely to both make errors in inputting availability (leading to more ERII-Add and ERII-Delete events) and override the algorithm's schedules to meet the unique demands of that period. Thus, the error term in our regression may be correlated with ERII-Add and ERII-Delete, leading to potential endogeneity issues.

To address this issue, we modify Equation 1 to include store $\times$ year-week fixed effects. The estimates of this table are presented in Column (6) of Table 1. We find that the effect sizes decrease, but retain their statistical significance at the 1% level. Specifically, a 1% increase in ERII-Add and ERII-Delete corresponds to a 1.2 and 0.25 percentage point increase in the probability of an override.

Next, we study the potential of non-linear effects, which were discussed in Hypothesis 2. To do so, we estimate Equation 1 after including quadratic terms for ERII-Add and ERII-Delete. We also retain the interacted store $\times$ year-week fixed effects. Column (7) of Table 2 presents these estimates. Here, we find a positive linear and quadratic effect for ERII-Add, but a positive linear, but negative quadratic effect for ERII-Delete. One explanation for this is the collinearity of ERII-Add and ERII-Delete. To determine the overall effect of ERII on overrides, we present the summed effect of ERII-Add and ERII-Delete in Column (8). Here, we find a positive linear and quadratic effect of ERII on overrides, which are significant at the 1% level. Accordingly, we find evidence in support of non-linear spillover effects, as discussed in Hypothesis 2.

6.3 Relationship with Algorithm Aversion

Overall, consistent with our hypothesis, we find that the presence of ERII-Add and ERII-Delete events in a store during a given week correlates with a higher likelihood of ex-post overrides being applied to shifts not directly affected by these events. These results are also consistent with the hypothesis that, to accommodate ERII-Add and ERII-Delete, managers might have to adjust other shifts that are not subject to these modifications. For example, consider a scenario where, due to an input inaccuracy, John's shift is manually added (ERII-Add) to a particular day. In order to incorporate John's shift, it may become necessary to adjust the shifts of other employees working that day, such as Sally and Stu, to ensure optimal staffing and avoid overlaps. Thus, despite Sally and Stu's shifts not being directly subjected to ERII-Add or ERII-Delete, they may be overridden to accommodate John's manually added shift. In other words, the ex-post overrides can "spillover" to shifts not directly linked to ERII-Add or ERII-Delete.

More broadly, our results provide empirical evidence that one factor underlying "algorithm aversion" (Balakrishnan et al., 2022; Dietvorst and Bharti, 2020; Dietvorst et al., 2015) may result from incorrect inputs. That is, in addition to the possibility that human operators have private information (over the AI tool) or conflicting opinions and judgements about the "correct" decision, our results suggest that overrides may frequently occur as a consequence of incorrect inputting of algorithm inputs. Accordingly, it is conceivable that, in a counterfactual world where the inputs are specified correctly, the scheduling output generated by the algorithm would be completely agreeable, requiring minimal or no overrides. Overall, our results underscores the crucial role of input accuracy in leveraging the full potential of AI tools, thereby contributing to a more nuanced understanding of the interplay between human judgement and algorithmic decision-making in organizational settings.

In this section below, we examine the direct and spillover effects of ERII-Add and ERII-Delete affect the "quality" of work schedules. We measure quality from the perspective of corporate officers and employees.

7 The Effects of ERII on Work Schedule Quality

To recap, in Section 4, we showed that ERII causes managers to make manual shift additions or deletions (direct effects). The previous section also revealed that ERII-subject shifts can also lead to overrides on non-ERII shifts (spillover effects). We now test the hypothesis that the direct and spillover effects of ERII may ultimately have a negative effect on the quality of work schedules that are distributed to employees (Hypothesis 3).

Recall from Section 5 that there are two potential mechanisms for this hypothesis. First, due to ERII, the set of schedules generated by the algorithm may be of poor quality, with managers not being able to fully rectify these the shortcomings stemming from ERII. Specifically, they may lack detailed insights and information possessed by the algorithm or be limited by time or in their judgement. Therefore, when ex-post overrides are made, they may be addressing immediate concerns but overlooking broader optimization goals or constraints. This mechanism relates to the direct effects of ERII.

Second, even if managers have some understanding of what work schedules should have looked like in the absence of ERII, the potentially large number of changes resulting from direct and spillover effects may introduce mistakes or mistakes: they may address one issue only to inadvertently create another. The frequent need for adjustments stemming from ERII may also reduce the manager's trust in the algorithm (despite the inaccurate inputs being the source of the problem), leading to further ad-hoc changes that deviate from the optimal work schedule. This mechanism relates to both the direct and indirect effects of ERII. An additional hypothesis that emerges based on this discussion is the potential non-linear effects of ERII: While managers may be able to rectify work schedules when they are subject to a relatively small number of ERII-subject shifts, the challenges in rectifying work schedules become increasingly more difficult.

7.1 Corporate Perspective of Work Schedule Quality

We define quality from the perspective of two relevant parties in our empirical context: corporate officers and employees. From the perspective of corporate officers, the quality of work schedules is synonymous with how closely labor is aligned with forecasted demand. We know this to be true based on conversations with corporate officers who utilize the underlying AI tool. Recall from Section 2 that the algorithm forecasts demand in 15-minute intervals of time, and the algorithm attempts to generate work schedules that exactly match the forecast. Corporate officers are directly involved in the forecasting process and in determining the levels of labor service. For example, they are involved in the selection process of the underlying inputs and paramaters (e.g., past inventory levels) that the algorithm uses to forecast demand. Consequently, from the perspective of corporate officers, labor that exceeds the demand forecast is considered wasteful (overage), and labor that falls below the demand forecast will lead to lost sales (underage). A negative association between ERII and misalignment with demand forecasts will be prima facie evidence for a reduction in work schedule quality from the perspective of corporate officers.

Based on this rationale, we proxy schedule quality from the perspective of corporate officer by capturing the total underage and overage of labor across all 15-minute intervals on a given day:

$$QualCorp_{i,t} = \frac{\sum_k (O_{i,t,k} + U_{i,t,k})}{\sum_k D_{i,t,k}} \tag{2}$$

where $O_{i,t,k}$ and $U_{i,t,k}$ capture the total overage and underage (in minutes) at store i on date t during labor interval k, and $D_{i,t,k}$ represents total demand (also in minutes). We are informed that corporate officers pay very close attention to the value of $QualCorp_{i,t}$ across their stores, and it is one of the first metrics that they examine in order to understand the performance differences in their stores. In our empirical analysis below, we examine whether ERII reduces the value of $QualCorp_{i,t}$ decreases between V1 and V2.

7.2 Employee Perspective of Work Schedule Quality

Next, from the perspective of workers, we define the quality of work schedules based on its *consistency* across weeks. A large literature in Operations Management and Sociology has documented the instability of employee work schedules, and their potential negative effects on employee well-being and performance (Ben-Ishai, 2015; Bergman et al., 2022; Harknett et al., 2021; Henly and Lambert, 2014; Kwon and Raman, 2023b,c; Lambert et al., 2019; Lu et al., 2022). One key takeaway from this body of papers is that inconsistent work schedules can lead to disruptions in employees' daily routines, difficulties in managing personal commitments, and potential financial hardships due to fluctuating hours. Such instability can also lead to increased stress and decreased job satisfaction. Furthermore, when workers can predict their hours, they can make informed decisions about childcare, continuing education, and other personal pursuits.

Three recent papers closely relate inconsistent work schedules with store operations. First, inconsistent work schedules can reduce the morale of employees, which may result in a lower effort (Lu et al., 2022). Second, it can increase employee turnover (Bergman et al., 2022). Finally, it can result in higher lateness and absenteeism rates of employees, which have clear associations with store performance (Kwon and Raman, 2023a) due to the shortfall of labor they may cause.

Based on this rationale, we proxy the quality of work schedules from the perspective of workers based on the consistency of their distributed work schedules. Our hypothesis is that, in attempting to rectify the errors caused by ERII, managers may make overrides that inadvertently increase the inconsistency in workers' schedules.

Specifically, managers may sometimes not remember or not factor in an employee's previous work schedule when making overrides, which may lead to increased inconsistency of work schedules. In contrast, the schedules distributed by the algorithm may have greater consistency not because the algorithm explicitly factors this in, but because the algorithm systematically adheres to the provided employees' available days and start times, ensuring a more regular pattern in scheduling. Consequently, due to the direct and spillover effects of ERII, we hypothesize a reduction in the consistency of work schedules for employees.

To test this lattter hypothesis, we follow Bergman et al. (2022); Kwon and Raman (2023a,b); Lu et al. (2022) and define four measures of schedule quality at the store-person-year-week levels:

1. The total number of shifts that are scheduled on a day of the week (e.g, Monday) that the employee did *not* work in the previous week (referred to as day-of-the-week inconsistency). We label this variable as $\tilde{N}_{i,t,j}$.
2. The total number of shifts that are scheduled on a day of the week that the employee *did* work on, but where the start time differs by more than one hour (referred to as start-time inconsistency). We label this variable as $N^*_{i,t,j}$.
3. The total number of shifts that are scheduled on back-to-back days but where the start time is different by more than one hour[10]. We label this variable as $\widehat{N}_{i,t,j}$, and refer to it as back-to-back start-time inconsistency..
4. Week-to-week fluctuations in total hours, which we define by:

$$\widehat{H}_{i,j,t} = (H_{i,j,j} - \bar{H}_{i,j})^2 \tag{3}$$

In the above, variables are defined at the, store (i), worker (j), and year-week (t) levels. In the above equation, $H_{i,j,t}$ represents the total number of hours scheduled, and $\bar{H}_{i,t,j}$ is the average amount of scheduled labor in hours for employee j at store i.

Overall, our analysis in this section examines the cumulative effects on the quality of work schedules resulting from the direct and spillover effects of ERII. Therefore, we study the effects of ERII on $QualCorp_{i,t}$, $H^*_{i,j,t}$, and $\widehat{H}_{i,j,t}$ on our complete sample of employees and store observations, regardless of whether the focal store or the employee was subject to ERII.

7.3 Effects on Work Schedule Quality (Corporate Perspective)

Like our approach in examining spillovers (Section 6.1), we estimate fixed effects models to study the effects of ERII on the quality of work schedules. First, Table 4 presents our results that study the effects of ERII on the quality of work schedules from the perspective of corporate officers, $QualCorp_{i,t}$. The dependent variable in this column is the difference in $QualCorp_{i,t}$ between V1 and V2, which we refer to as $\Delta QualCorp_{i,t}$. Therefore, these regressions test whether the presence of ERII-Add and ERII-Delete cause a degradation in the alignment with demand forecasts due to the overrides that managers make. We regress $\Delta QualCorp_{i,t}$ on ERII-Add and

[10]Note that we only count the second shift in our count. For instance, if a shift began on Monday at 9AM, but at 12PM on Tuesday, this would only be one instance of this form of start-time inconsistency.

ERII-Delete (among other control variables and fixed effects) on a sample of more than 7.7 million store-week observations.

Table 4 presents our results on the effects of ERII-Add and ERII-Delete on the quality of work schedules. Our preferred specification (Column 5) that includes all fixed effects and control variables, we find that a 1% increase in the number of ERII-Add subject shifts and ERII-Delete subject shifts lead to a 1.9 and 2.4 percentage point decrease in the alignment of work schedules between Versions 1 and 2, respectively. These estimates are all statistically significant at the 1% level (standard errors are double clustered at the store and year-week levels). Based on the mean value of $QualCorp_{i,t}$ in V1 of 0.674, these estimated effects are relatively large. Accordingly, we find that the ERII causes a degradation in the quality of work schedules from the perspective of corporate officers. These effects may also translate to a reduction in store performance if the algorithm's forecasts are reliable.

Due to our included fixed effects (store and year $\times$ week), we note that our results from our preferred specification are not driven by unobservables that do not vary within stores and within year-week groups of observations. Key unobservables that are controlled for by the store fixed effect include store size and geographical characteristics associated with the store. The size of the store is important simply because the size of the store is correlated with the size of the staffing roster. Furthermore, it may be the case that the AI tool finds it more challenging to forecast demand for AI tools. Accordingly, without store fixed effects, the effect of ERII may be biased towards a large effect because it captures the fact that the AI tool may generate less accurate forecasts for larger stores, on average. In addition, year $\times$ week fixed effects account for seasonal factors that may affect demand and turnover (which may amplify issues with ERII). Finally, our three control variables (scheduled labor in V1, operating duration, and demand forecast) control for demand-related factors that may vary within groups of store and year $\times$ week observations.

However, despite these fixed effects and controls, there are additional concerns with omitted variables that may bias our estimates on the effects of ERII on work schedule quality (from the perspective of corporate officers). A key concern relates to the ability of the store manager. "Good" store managers may have fewer instances of ERII, and may also be better able to rectify issues associated with work schedules that are subject to ERII. Specifically, it may be the case that these "good" store managers are more meticulous with how they schedule labor in their stores, which includes being more aware of the inputs that are fed into the AI tool. Furthermore, it is possible that even during instances where ERII occurs due to forgetfulness or oversight, these "good" managers may have a better understanding of the operational needs of their stores, allowing their stores to be less affected by ERII. However, due to the mobility of managers across stores, our current set of fixed effects are insufficient to address this potential source of bias.

To address the bias associated with managerial ability (among other time-varying unobservables), we interact store $\times$ year-week fixed effects. In doing so, we estimate the effects of ERII only from within store $\times$ year-week groups of observations, which account for the fact that management may change within the cross-section of our stores. Column (6) of Table 4 presents these results. We find that the effect of ERII-Add on $\Delta QualCorp_{i,t}$

is identical (-1. 9%), with the same level of significant at the 1% level. In contrast, we find that the effect of ERII-Delete decreases considerably from 2.4% to 0.2%. However, the overall message remains the same: the presence of ERII decreases work schedule quality from the perspective of corporate officers, as managers make overrides to decrease the alignment of labor from the demand forecasts and service levels that are set by the AI tool and corporate officers.

We now study examine the potential non-linear effects of ERII on work schedule quality. As suggested above, managers may be able to effectively rectify work schedules that are subject to a few instances of ERII. However, as the number of ERII-subject shifts increases, the complexity of making appropriate adjustments increases nonlinearly, which potentially amplifies the negative effects of ERII on work schedule quality from the perspective of corporate officers. The evidence to support this hypothesis is found in Column (7) of Table 4. Specifically, we find that the linear and quadratic terms for ERII-Add are negative and statistically significant at the 1% level, which supports the *cascading effect* that we discussed when presenting Hypothesis 3, which posited a relationship between ERII and the quality of the work schedule. In contrast, while both the linear and quadratic effect are negative for ERII-Delete, only the linear effect is statistically significant (ta the 1% level).

7.4 Effects on Work Schedule Quality (Employee Perspective)

Next, we examine the effects of ERII on the quality of work schedules from the perspective of employees. Recall from Section 7.2 that we constructed three proxies of schedule quality from the perspective of workers. These measures capture inconsistency with respect to: (1) the scheduled day-of-the week (V_i^1, t, j), (2) start time ($V_{i,t,j}^2$), and (3) scheduled hours for workers ($\widehat{H}_{i,t,j}, H^*_{i,t,j}$) inconsistency. To examine these effects we regress these variables on ERII-Add and ERII-Delete store, person, year$\times$ week fixed effects. We also include the same control variables (at the store-date level) used in our previous analysis, where studied the effects of ERII on the quality of work schedules from the perspective of corporate officers.

Table 5 presents the regression estimates of our measures of work schedule quality from the perspective of workers on ERII-Add and ERII-Delete. These estimates are based on a sample of more than 13 million store-worker-year-week groups of observations. Each column uses one of the four dependent variables that capture the the inconsistency of the work schedule discussed in Section 7.2. Consistent with our hypothesis that the algorithm's set of generated work schedules is more systematic and consistent, we find evidence that the edits caused by ERII decrease work schedule consistency for workers. Specifically, we find that a 1% increase in ERII-Add results in a 1.0% , 1.6% and 2.8% increase in $\hat{N}$, $\tilde{N}$ and N^*, respectively. Moreover, a 1% increase in ERII-Delete reults in a 0.7%, 0.7%, and -0.03% increase in the same three variables. All point estimates are significant at the 1% level when double clustering standard errors at the store- employee and year-week levels. We note that one of the point estimates is negative (associated with ERII-Delete and N^*). This may result from the collinearity of ERII-Add and ERII-Delete. However, when examining the total effect of ERII (i.e., the sum of ERII-Add and ERII-Delete), we find a positive effect, where a 1% increase in the total number of shifts subject to ERII corresponds to a 1.3% increase in N^*.

7.5 The Potential Effects on Store Performance

Overall, we find in the previous two sections that ERII can degrade work schedule quality, where quality is proxied from both the perspective of corporate officers and employees. We have two hypothesized mechanisms for this result. First, in the pursuit of making ex-post overrides that stem from ERII, managers steer their work schedules further from being aligned with demand forecasts. This may occur due to the practical challenges of moving labor around in a way that fixes ERII *and* still adheres to demand forecasts. In addition, in cases where the magnitude of ERII-Delete is large, managers may simply not have the staffing roster to schedule workers in a way that aligns with the demand forecast. Ultimately, if these demand forecasts are correct on average, it is possible that this degradation in the quality of the work schedule will also have a negative effect on store performance.

Second, we found that ERII also caused work schedules to become more inconsistent for workers. One explanation for this result is that, in the pursuit of rectifying ERII-subject work schedules, managers make overrides in a way that does not account for employees' previous work schedules. This may occur simply because managers are not able to fully remember when and where their employees worked in previous weeks. Based on recent empirical studies on the effects of inconsistent work schedules on productivity, lateness and absenteeism, and employee turnover (Bergman et al., 2022; Kwon and Raman, 2023a; Lu et al., 2022), the degradation of the quality of work schedules from the perspective of employees may also ultimately have a negative effect on store performance.

Finally, an additional channel through which ERII can degrade store performance is the time and effort lost by managers to correct ERII-related issues. Correcting ERII may consume managerial resources that could be better spent on other critical aspects of store management, such as staff training, customer relations, or inventory management. This diversion of managerial attention away from other important tasks can further have a negative effect on store performance.

8 Conclusion and Discussion

Firms are commonly reported to have no significant financial returns from AI tools (Ransbotham et al., 2020). The literature has focused on the possibility that human operators are using AI tools suboptimally. Specifically, there is some evidence that the overrides to the AI tool made by human operators negatively affects operational performance. For example, Kesavan and Kushwaha (2020) find that overrides made to a decision support system that selects products have negative effects on store profit. However, there is also evidence that overrides can also have positive effects on operational performance if managers have private knowledge that cannot easily and systematically be incorporated as inputs into the AI tool (Kwon et al., 2023).

In this paper, we provide an alternate explanation for the limited financial returns of AI tools. Specifically, we show that human operators may be incorrectly specifying key inputs that are used by AI tools. In the context of algorithmic labor scheduling, we show that store managers frequently and persistently misspecify the availability of their employees. While this issue may appear strongly hypothetical, we provide evidence for this phenomenon by examining data from four independent retail chains that use AI tools to automate their labor scheduling decisions.

We first show that because of incorrect input records that specify the availability of every employee, the AI tool generates erroneous work schedules. We refer to this problem as Employee Record Input Inaccuracy ("ERII"), and we identify two major cases. First, employees that were no longer active at a particular store were scheduled by the AI tool because the store manager did not remove them from the set of available employees. As a result, the store manager had to manually remove them from the set of generated work schedules. Second, employees who were available to work at a particular store did not receive any work schedules. The latter case was identified by observing work schedules that were ultimately distributed to employees who had no availability records. The AI tool can never generate work schedules for employees without an availability record.

Beyond the direct effects of having to rectify erroneous work schedules, we also find that ERII has spillover effects on shifts that were *not* subject to ERII. Using a knitted sweater as an analogy, our explanation for these spillover effects is that due to the interconnected nature of work schedules (akin to threads in a sweater), making an adjustment to a single shift that is subject to ERII may require adjustments to other shifts. For example, a shift that needs to be manually deleted because it was assigned to an employee that has left the store may require the manager to bring in another employee to cover the vacated shift. However, this new assignment could conflict with that employee's previously assigned responsibilities on a different day or time. As a result, the manager then has to reshuffle multiple shifts, reallocating resources, and potentially inconveniencing several employees in the process.

We find that ERII is not merely a potential nuisance that takes up management time in readjusting schedules. Specifically, we find that ERII negatively affects the quality of work schedules from both the perspective of corporate officers and employees. Specifically, ERII causes further misalignment between scheduled labor and the required amount of labor that is determined by the AI tool after it makes demand forecasts. We also find that ERII causes work schedules to become more inconsistent for workers. Overall, we find that ERII can impose significant time opportunity costs as managers attempt to rectify work schedules that are subject to ERII. We also find that ERII can have negative effects on two proxies of work schedule quality, which have plausible links to store performance.

We conclude by commenting on the generalizability of our findings. First, we note that ERII is a problem for multiple independent retailers and their stores, which does not suggest that it is an isolated problem within a single organization. Second, we note that while our empirical analysis focuses on a singular operational context of labor scheduling, it may be possible that issues of input inaccuracy are also present when AI tools are used to automate or improve up on core operational tasks. For example, AI tools that forecast demand hinge upon the volume and accuracy of past sales data. Inaccurate input records can occur, for example, if the operator of the AI tool mixes sales from one department with another (e.g., shoes versus handbags). While this example may seem strongly hypothetical, it is worth emphasizing that input inaccuracies in our context of scheduling labor appeared just as hypothetical. Indeed, store managers must have noticed this problem when manually removing employees who had left the store. However, we found that input inaccuracies were large, persistent, and went largely unnoticed by corporate officers of the retailers underlying our study. Consequently, a key managerial takeawya from our paper is that corporate officers may benefit from establishing clear guidelinse and offering training protocols to ensure that managers are accurately inputting data when using AI tools.

References

Balakrishnan, M., Ferreira, K., and Tong, J. (2022). Improving human-algorithm collaboration: Causes and mitigation of over- and under-adherence. *Working Paper*.

Ben-Ishai, L. (2015). Volatile job schedules and access to public benefit. *CLASP*.

Bergman, A., David, G., and Song, H. (2022). I quit: Schedule volatility as a driver of voluntary employee turnover. *Working Paper*.

Brynjolfsson, E. and McElheran, K. (2019). Data in action: Data-driven decision making and predictive analytics in u.s. manufacturing. *Working Paper*.

Byrne, M. F., Parsa, N., Greenhill, A. T., Chahal, D., Ahmad, O., and Bagci, U., editors (2023). *AI in Clinical Medicine: A Practical Guide for Healthcare Professionals*. Wiley-Blackwell.

Caro, F. and Tejada Cuenca, A. S. d. (2018). Believing in analytics: Managers' adherence to price recommendations from a dss. *Working Paper*.

Cui, Y., Jiang, Z. Z., and Li, Q. (2023). Ai assistance and service usage. Working paper.

Dai, T. and Abramoff, M. (2023). Incorporating artificial intelligence into healthcare workflows: Models and insights. *INFORMS TutORials in Operations Research*.

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

DeHoratius, N. and Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science*, 54(4):627–641.

Dietvorst, B. J. and 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, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114–126.

DiSorbo, M. D. and Ferreira, K. (2023). Warnings and endorsements: Improving human-algorithm collaboration under dataset shift. Working paper.

Harknett, K., Schneider, D., and Irwin, V. (2021). Improving health and economic security by reducing work schedule uncertainty. *Proceedings of the National Academy of Sciences*, 118(42):e2107828118.

Henly, J. R. and Lambert, S. J. (2014). Unpredictable work timing in retail jobs: Implications for employee work–life conflict. *ILR Review*, 67(3):986–1016.

Hoffman, M., Kahn, L. B., and Li, D. (2017). Discretion in Hiring. *The Quarterly Journal of Economics*, 133(2):765–800.

Huang, M.-H. and Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, 21(2):155–172.

Ibanez, M. R., Clark, J. R., Huckman, R. S., and Staats, B. R. (2017). Discretionary task ordering: Queue management in radiological services. *Management Science*.

Kawaguchi, K. (2020). When will workers follow an algorithm? a field experiment with a retail business. *Management 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. and 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.

Kwon, C. and Raman, A. (2023a). The effects of inconsistent work schedules on employee lateness and absenteeism. *Working Paper*.

Kwon, C. and Raman, A. (2023b). Understanding the effects of fair workweek laws on worker schedules. *Working Paper*.

Kwon, C. and Raman, A. (2023c). The value of flexible scheduling practices: Evidence from fair workweek laws. *Working Paper*.

Kwon, C., Raman, A., and Tamayo, J. (2023). Human-computer interactions in demand forecasting and labor scheduling decisions. *Working Paper*.

Lambert, S. J., Henly, J. R., and Kim, J. (2019). Precarious work schedules as a source of economic insecurity and institutional distrust. 5(4):218–257.

Lebovitz, S., Levina, N., and Lifshitz-Assaf, H. (2021). Is ai ground truth really true? the dangers of training and evaluating ai tools based on experts' know-what. *Management Information Systems Quarterly*, 45(3).

Lin, W., Kim, S.-H., and Tong, J. (2022). What drives algorithm use? an empirical analysis of algorithm use in type 1 diabetes self-management. Available at SSRN 3891832.

Liu, X., Susarla, A., and Padman, R. (2022). Ai for social good: A human-in-the-loop augmented intelligence video understandability assessment to promote health literacy. Working paper.

Lu, G., Du, R. Y., and Peng, X. D. (2022). The impact of schedule consistency on shift worker productivity: An empirical investigation. *Manufacturing & Service Operations Management*, 24(5):2780–2796.

Netessine, S., Fisher, M., and Krishnan, J. (2010). Labor planning, execution, and retail store performance: An exploratory investigation. *Working Paper*.

Ransbotham, S., Khodabandeh, S., Kiron, D., Candelon, F., Chu, M., and LaFountain, B. (2020). Expanding ai's impact with organizational learning. Technical report, MIT Sloan Management Review and Boston Consulting Group.

Snyder, C., Keppler, S., and Leider, S. (2022). Algorithm reliance under pressure: The effect of customer load on service workers. *Working Paper*.

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

van Donselaar, K. H., Gaur, V., van Woensel, T., Broekmeulen, R. A. C. M., and Fransoo, J. C. (2010). Ordering behavior in retail stores and implications for automated replenishment. *Management Science*, 56(5):766–784.

A Supplementary Appendix

A.1 Examples of ERII

We provide real examples of ERII-Add and ERII-Delete in Tables 6 and 7, respectively. These examples help illustrate their nature in the data. Table 6 presents three potential instances of ERII-Add in the data. In Panel A, the employee, John, is hired on June 1, 2021, and begins work on June 6, 2021. However, due to management not having set his availability correctly, we observe that the algorithm does not generate shifts for John, but they are manually generated and distributed by the manager for consecutive weeks. In Panel B, we see that Kyle's work schedules were successfully generated approximately two weeks after his hire date (more on this timing in the following example). After having worked only one week, Kyle requested two weeks off (as revealed in the HR data), and based on the manager having approved the request, the algorithm did not schedule him in the fourth week. However, we see that his return was not correctly specified, and while he resumed work after two weeks, his work schedule was manually generated and distributed by his manager.

The final example of ERII-Add (presented in Panel C of 6) shows a potential example of ERII-Add that requires some discretion. In this example, Heidi is hired on June 1, 2021 and has her availability correctly specified and inputted into the algorithm (which we observe). However, the manager would need to manually construct her work schedules, as Heidi missed the scheduling window for the upcoming 2-3 weeks[11]. The same phenomena also occurs for Kyle (Panel B), as he begins work two weeks before his first work schedule is generated by the algorithm.

We do not consider these instances as a ERII-Add because they do not indicate a scenario where the algorithm fails to schedule an employee due to inaccuracies in the employee record inputs. Here, the issue is not with the accuracy of the input data. Rather, the challenges arise from the operational mechanics of the scheduling algorithm, which schedules 2-3 weeks in advance, and does not accommodate newly hired employees who miss this scheduling window. Therefore, the need for manual intervention in these cases is not a consequence of inaccurate employee records, but a result of the algorithm's design and operation. Consequently, to make our measure conservative, we do not classify manual additions of entire work schedules in the first *three* weeks of an employee's start date at a store as ERII-Add.

This filter is conservative for two reasons. First, there are many instances where the employee is hired, but wait to work until their schedules are generated by the algorithm. Second, there are many instances where the burn period captures consecutive instances of ERII-Add that persist beyond the initial three weeks. This indicates that the employee's availability was not correctly specified at the time of hiring. However, the burn period makes the ERII-Add magnitude appear smaller than it is in reality. Finally, we note that we also have a similar "burn"

[11] See Section 2.2 for additional details. However, in short, because the algorithm schedules 2-3 weeks in advance, a new hire can only have their schedules generated 2-3 weeks later at the earliest (if their availability is inputted immediately after their hire). Recall that stores can vary in how far in advance the algorithm schedules labor.

period for ERII-Delete, which we describe after presenting examples of ERII-Delete below.

Next, Table 7 presents three potential examples of ERII-Delete. In Panel A, Anna had a temporary absence on the week of July 5 2021, that was correctly specified by the manager (no schedules were generated by the algorithm, and the manager took no further action). However, the manager incorrectly specified Anna's return to be on the following week (July 12), when in reality, she was able to work two weeks after (July 19). Consequently, the manager had to manually delete her shifts on the Week of July 12, 2021. Next, Panel B presents an example of ERII-Delete stemming from a manager not removing the employee (Bob) as an input despite the employee having left the store (indicated by HR records). Here, we see that the algorithm generates work schedules for four weeks after Bob left the store, which the manager had to manually remove. The final example in Panel C shows a potential example of ERII-Delete that mirrors a similar concern to instances of ERII-Add occurring mechanically due to the algorithm scheduling 2-3 weeks in advance. Specifically, we see two instances of ERII-Delete for Charlie in the final two weeks of employment. However, we do not categorize these instances as ERII-Delete. Our primary concern lies in the ambiguity surrounding whether Charlie provided the standard two weeks notice to their manager (of him quitting) or not, and whether, if given, the manager failed to update the algorithm accordingly. To be conservative in our analysis, we exclude from consideration any instances of ERII-Delete occurring in the last three weeks of an employee's tenure at a given store. Accordingly, the "burn period" of ERII-Add and ERII-Delete are symmetric, and set at 3 weeks.

We conclude this subsection by noting that our definitions of inaccurate employee record input (for both ERII-Add and ERII-Delete) is a lower bound (and hence, conservative) because it only captures *across*-week differences in the actual availability of workers and the availability that the algorithm conditions on. Consequently, it does not capture *within*-week differences in employee availability, such as availability or the inability to work certain days of the week or during certain periods of the day. For example, managers may incorrectly input into the algorithm that a worker is available on Monday, Wednesday, and Friday, when they are only actually available to work on Tuesdays and Thursdays. Similarly, managers may input a start time availability of 9AM for a worker, when in reality they are only available from 11AM. Managers may make edits to work schedules that reflect the true availability of workers within a week by making overrides. However, it is difficult to know from the data alone whether these overrides are due to considerations of employee availability rather than differences in expectations and beliefs between the algorithm and the manager. For example, a manager may re-allocate an employee from the morning to the afternoon, say, but our data cannot distinguish whether this change is based on considerations of the employees' availability or if the manager believes that the afternoon will have higher customer demand than in the morning.

Employee's Availability Records Correctly Set

Employee Available to Work	Yes	No
Yes	No Issues (Employee Successfully Scheduled)	ERII-Add (Employee Manually Added)
No	ERII-Delete (Employee Manually Deleted)	No Issues (Employee left store/company)

Figure 1: Four Scenarios of Algorithmic Work Schedule Generation Based on the Alignment or Misalignment of HR Records and Employee Availability Records

Notes: The above diagram illustrates the four scenarios of work schedules that are possible based on whether Employee Availability Records are consistent. Both tables, as well as the term "consistency", are introduced and discussed in Section 2. Here, we describe the four scenarios beginning from the top left quadrant, and proceeding clockwise. First, (Yes,Yes) indicates a scenario where an employee is employed at a given store, and his or her availability is input into the algorithm. No issues with with algorithmic work schedule generation occur here. Next, (Yes, No) indicates a scenario where the employee is employed at a given store, but his or her availability is *not* input into the algorithm. Here, the manager must manually add the employee to the set of distributed work schedules. (No, No) indicates a scenario where the employee has left the store, but their availability was correctly updated to being not available. No issues with algorithmic work schedule generation occur here, because a schedule will never be generated. Finally, (No,Yes) indicates a scenario where the employee has left the store, which is indicated by not being present in HR records. However, their availability records have not yet been updated, and work schedules are potentially generated for that employee.

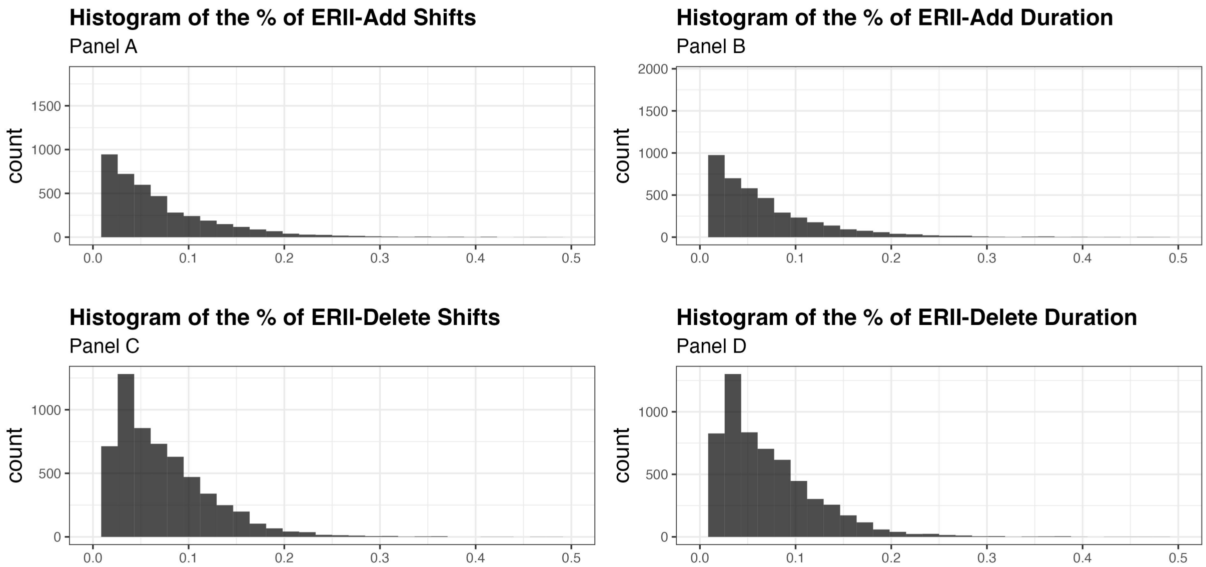

Figure 2: Timing of Overrides

Notes: The above histograms are constructed from the sample containing cross-sectional store data (See Panel C of Table 1). Therefore, each observation in the histogram is a store. Each panle differs based on the measure of ERII. Panel A is the ratio between ERII-Add Shifts and total V2 shifts, Panel B is the ratio between ERII-Add Duration and total V2 minutes, Panel C is the ratio between ERII-Delete Shifts and total V2 shifts, and Panel D is the ratio between ERII-Delete duration and total V2 minutes.

Table 1: Summary Statistics

Variable	N	Mean	SD	Median
Panel A: Shift-Level Outcomes				
V1 Labor Minutes	48,243,780	430.3774	104.673	480
V2 Labor Minutes	51,348,726	432.8855	101.329	480
Only V1 Exists	74,329,783	0.3092	0.462	0
V1 and V2 Exists	74,329,783	0.3399	0.474	0
Only V2 Exists	74,329,783	0.3509	0.477	0
Override Made to Shift (Yes = 1)	74,329,783	0.8366	0.37	1
V1 Break Duration	48,243,780	28.774	30.278	30
V2 Break Duration	51,348,726	18.1759	21.29	15
ERII-Add	74,329,783	0.0496	0.217	0
ERII-Delete	74,329,783	0.0337	0.181	0
ERII-Add or Delete	74,329,783	0.0834	0.276	0
Panel B: Store-Date Level Outcomes				
V1 Labor Minutes	7,722,410	2679.0429	3551.384	1245
V2 Labor Minutes	7,722,410	2863.0308	4157.606	1170
V1 Shifts	7,722,410	6.2252	8.172	3
V2 Shifts	7,722,410	6.6135	9.372	3
Operating Duration in Minutes	7,722,410	716.5797	346.272	600
Demand Forecast Converted to Labor Minutes	7,722,410	2904.4928	3746.117	1170
V2 - V1 Demand Alignment	7,722,410	-0.1947	0.286	-0.1629
ERII-Add (Shifts)	7,722,410	0.4712	1.759	0
ERII-Delete (Shifts)	7,722,410	0.3207	0.693	0
ERII-Add (Duration)	7,722,410	201.0465	726.078	0
ERII-Delete (Duration)	7,722,410	131.5799	294.637	0
ERII-Add (Shifts) / V2 Shifts	7,181,253	0.0639	0.189	0
ERII-Delete (Shifts) / V2 Shifts	7,181,253	0.0714	0.171	0
ERII-Add (Duration) / V2 Minutes	7,722,410	0.0583	0.18	0
ERII-Delete (Duration) / V2 Minutes	7,722,410	0.0679	0.175	0
Panel C: Store Level Outcomes (Cross-Sectional)				
ERII-Add (Shifts)	5,978	617.3108	1312.819	115
ERII-Delete (Shifts)	5,978	419.2451	413.734	296
ERII-Add (Duration)	5,978	263421.7297	567012.027	42285
ERII-Delete (Duration)	5,978	172066.7238	176151.109	117990
V1 Shifts	5,978	8070.2131	10292.625	3729
V2 Shifts	5,978	8589.6163	11755.908	3612.5
V1 Labor Minutes	5,978	3473237.6146	4503271.198	1567942.5
V2 Labor Minutes	5,978	3718320.0376	5231652.799	1467675
ERII-Add (Shifts) / V2 Shifts	5,978	0.0584	0.099	0.0298
ERII-Delete (Shifts) / V2 Shifts	5,978	0.0839	0.112	0.061
ERII-Add (Duration) / V2 Minutes	5,978	0.0565	0.098	0.0279
ERII-Delete (Duration) / V2 Minutes	5,978	0.083	0.121	0.0582

Notes: The above table presents summary statistics at the shift-level (Panel A), store-date level (Panel B), and at the store-level (Panel C). These summary statistics cover all retailers and all their stores. The data's time period is between 2018 and 2023. However, as we have described in Section 2, the longitudinal data in Panels B and C are unbalanced because the data from companies start at different periods, and due to the entry and exit of stores.

Table 2: Effects of ERII on Ex-Post Overrides

	Override							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(1 + ERII-Add Shifts)	0.018***		0.017***	0.021***	0.019***	0.012***	0.006***	
	(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0002)	(0.0004)	
log(1 + ERII-Delete Shifts)		0.009***	0.008***	0.020***	0.013***	0.002***	0.005***	
		(0.0003)	(0.0003)	(0.0004)	(0.0003)	(0.0002)	(0.0004)	
log(1 + ERII-Add Shifts)^2							0.005***	
							(0.0003)	
log(1 + ERII-Delete Shifts)^2							-0.002***	
							(0.0004)	
log(1 + ERII-Total Shifts)								0.005***
								(0.0004)
log(1 + ERII-Total Shifts)^2								0.003***
								(0.0002)
Constant				1.93***				
				(0.003)				
log(Operating Mins.)				-0.120***	0.022***	-0.004***	-0.004***	-0.004***
				(0.0007)	(0.0009)	(0.0005)	(0.0005)	(0.0005)
log(Demand Forecast Mins.)				0.044***	0.002***	0.003***	0.002***	0.003***
				(0.0009)	(0.0007)	(0.0003)	(0.0003)	(0.0003)
log(V1 Mins.)				-0.083***	-0.059***	-0.033***	-0.033***	-0.035***
				(0.0009)	(0.001)	(0.0003)	(0.0003)	(0.0003)
Observations	69,453,609	69,453,609	69,453,609	69,135,432	69,135,432	69,135,432	69,135,432	69,135,432
Adjusted R^2	0.101	0.100	0.101	0.047	0.102	0.201	0.201	0.201
Store FEs	✓	✓	✓		✓			
Year-Week FEs	✓	✓	✓		✓			
Store-Year-Week FEs						✓	✓	✓

Notes: The above table presents fixed effects regression estimates associated with Equation 1 on the sample of shifts that are *not* subject to ERII as defined in Section 3.2. The dependent variable is a dummy variable that equals 1 if the shift experienced an override as defined in Section 2. That is, if the shift was either manually added, manually deleted, or had any of the five shift characteristics modified between V1 and V2. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered at the store and year-week levels.

Table 3: Effects of ERII on Schedule Alignment with Demand Forecast

	Diff. Schedule Alignment (V2-V1)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log(1 + ERII-Add Shifts)	-0.021***		-0.020***	-0.038***	-0.019***	-0.019***	-0.010***
	(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0003)	(0.0006)
log(1 + ERII-Add Shifts)^2							-0.008***
							(0.0004)
log(1 + ERII-Delete Shifts)		-0.015***	-0.014***	-0.048***	-0.024***	-0.002***	-0.002***
		(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0003)	(0.0007)
log(1 + ERII-Delete Shifts)^2							-0.0002
							(0.0007)
Constant				0.342***			
				(0.003)			
log(V1 Mins.)				0.094***	0.098***	0.147***	0.147***
				(0.0007)	(0.0009)	(0.0006)	(0.0006)
log(Operating Mins.)				-0.063***	0.057***	-0.022***	-0.022***
				(0.0007)	(0.0008)	(0.0006)	(0.0006)
log(Demand Forecast Mins.)				-0.108***	-0.085***	-0.095***	-0.095***
				(0.0008)	(0.0008)	(0.0006)	(0.0006)
Observations	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410
Adjusted R^2	0.298	0.297	0.298	0.050	0.311	0.645	0.645
Store FEs	✓	✓	✓		✓		
Year-Week FEs	✓	✓	✓		✓		
Store-Year-Week FEs						✓	✓

Notes: The above table presents fixed effects regression estimates estimated at the store-date level. The dependent variable is the difference in schedule alignment (Equation 2) between Versions 1 and 2. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered at the store and year-week levels.

Table 4: Effects of ERII on Schedule Alignment with Demand Forecast

	Diff. Schedule Alignment (V2-V1)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
log(1 + ERII-Add Shifts)	-0.021***		-0.020***	-0.038***	-0.019***	-0.019***	-0.010***
	(0.0004)		(0.0004)	(0.0004)	(0.0004)	(0.0003)	(0.0006)
log(1 + ERII-Add Shifts)^2							-0.008***
							(0.0004)
log(1 + ERII-Delete Shifts)		-0.015***	-0.014***	-0.048***	-0.024***	-0.002***	-0.002***
		(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0003)	(0.0007)
log(1 + ERII-Delete Shifts)^2							-0.0002
							(0.0007)
Constant				0.342***			
				(0.003)			
log(V1 Mins.)				0.094***	0.098***	0.147***	0.147***
				(0.0007)	(0.0009)	(0.0006)	(0.0006)
log(Operating Mins.)				-0.063***	0.057***	-0.022***	-0.022***
				(0.0007)	(0.0008)	(0.0006)	(0.0006)
log(Demand Forecast Mins.)				-0.108***	-0.085***	-0.095***	-0.095***
				(0.0008)	(0.0008)	(0.0006)	(0.0006)
Observations	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410	7,722,410
Adjusted R^2	0.298	0.297	0.298	0.050	0.311	0.645	0.645
Store FEs	✓	✓	✓		✓		
Year-Week FEs	✓	✓	✓		✓		
Store-Year-Week FEs						✓	✓

Notes: The above table presents fixed effects regression estimates estimated at the store-date level. The dependent variable is the difference in schedule alignment (Equation 2) between Versions 1 and 2. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered at the store and year-week levels.

Table 5: Effects of ERII on Schedule Consistency

	$\log(1+\hat{N})$	$\log(1+\tilde{N})$	$\log(1+N^*)$	$\log(\hat{H}_{i,t})$	$\log(1+\hat{N})$	$\log(1+\tilde{N})$	$\log(1+N^*)$	$\log(\hat{H}_{i,t})$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
log(1 + ERII-Add Shifts)	0.010***	0.016***	0.028***	0.003***				
	(0.0002)	(0.0002)	(0.0002)	(0.0008)				
log(1 + ERII-Delete Shifts)	0.007***	0.007***	-0.003***	0.042***				
	(0.0001)	(0.0001)	(0.0001)	(0.0007)				
log(1 + ERII-Total Shifts)					0.012***	0.015***	0.013***	0.040***
					(0.0002)	(0.0002)	(0.0002)	(0.0008)
Observations	13,100,207	13,100,207	13,100,207	13,035,320	13,100,207	13,100,207	13,100,207	13,035,320
Adjusted R^2	0.350	0.293	0.220	0.267	0.350	0.293	0.218	0.266
Store-Employee FEs	✓	✓	✓	✓	✓	✓	✓	✓
Year-Week FEs	✓	✓	✓	✓	✓	✓	✓	✓

Notes: The above table presents fixed effects regression estimates estimated at the store-worker-week level. The dependent variables are schedule consistent metrics discussed in Section 7.2. Specifically, $\hat{N}$, $\tilde{N}$, N^*, $\hat{H}_{i,t}$ represent day-of-the-week inconsistency start-time inconsistency, back-to-back start time inconsistency, and weekly fluctuations in hours. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively. Standard errors are double clustered at the store and year-week levels.

Table 6: Examples of ERII-Add

Panel A: Employee's Availability Record Never Inputted				
Employee	Hire Date	Week Beginning	Algorithm Shifts	Distributed Shifts
John	2021-06-01	2021-06-06	0	4
John	2021-06-01	2021-06-14	0	4
John	2021-06-01	2021-06-21	0	4
John	2021-06-01	2021-06-28	0	4
John	2021-06-01	2021-07-05	0	3
John	2021-06-01	2021-07-12	0	4
⋮	⋮	⋮	⋮	⋮
Panel B: Employees' Availability Record Delay Due to Time-Off Return				
Employee	Hire Date	Week Beginning	Algorithm Shifts	Distributed Shifts
Kyle	2021-06-01	2021-06-06	0	4
Kyle	2021-06-01	2021-06-14	0	4
Kyle	2021-06-01	2021-06-21	4	4
Kyle	2021-06-01	2021-06-28	0	0
Kyle	2021-06-01	2021-07-05	0	0
Kyle	2021-06-01	2021-07-12	0	4
Kyle	2021-06-01	2021-07-19	0	4
⋮	⋮	⋮	⋮	⋮
Panel C: Employees' Availability Record Delay Due to New Hires/Store Transfers				
Employee	Hire Date	Week Beginning	Algorithm Shifts	Distributed Shifts
Heidi	2021-06-01	2021-06-06	0	4
Heidi	2021-06-01	2021-06-14	0	4
Heidi	2021-06-01	2021-06-21	4	4
Heidi	2021-06-01	2021-06-28	3	4
Heidi	2021-06-01	2021-07-05	4	3
Heidi	2021-06-01	2021-07-12	4	4
⋮	⋮	⋮	⋮	⋮

Notes: The statistics above present summary statistics for our entire sample. The estimates in Panels B, C, D, and E are constructed from our shift-level dataset (Panel A). Note that we intentionally dropped the number of observations in Panel C (cross-sectional summary statistics at the store-level to mask the number of stores in our sample.

Table 7: Examples of ERII-Delete

Panel A: Temporary Absences				
Employee	Termination Date	Week Beginning	Algorithm Shifts	Distributed Shifts
⋮	⋮	⋮	⋮	⋮
Anna	N/A	2021-06-21	4	4
Anna	N/A	2021-06-28	4	4
Anna	N/A	2021-07-05	0	0
Anna	N/A	2021-07-12	4	0
Anna	N/A	2021-07-19	4	4
⋮	⋮	⋮	⋮	⋮
Panel B: Never Removed				
Employee	Termination Date	Week Beginning	Algorithm Shifts	Distributed Shifts
⋮	⋮	⋮	⋮	⋮
Bob	N/A	2021-06-14	4	4
Bob	2021-06-21	2021-06-21	4	0
Bob	2021-06-21	2021-06-28	3	0
Bob	2021-06-21	2021-07-05	4	0
Bob	2021-06-21	2021-07-12	4	0
Panel C: Sudden Turnover				
Employee	Hire Date	Week Beginning	Algorithm Shifts	Distributed Shifts
⋮	⋮	⋮	⋮	⋮
Charlie	N/A	2021-06-14	4	4
Charlie	N/A	2021-06-21	4	4
Charlie	N/A	2021-06-28	3	4
Charlie	N/A	2021-07-05	4	3
Charlie	2021-07-12	2021-07-12	4	0
Charlie	2021-07-12	2021-07-19	4	0

Notes: The statistics above present summary statistics for our entire sample. The estimates in Panels B, C, D, and E are constructed from our shift-level dataset (Panel A). Note that we intentionally dropped the number of observations in Panel C (cross-sectional summary statistics at the store-level to mask the number of stores in our sample.