The Implications of Artificial Intelligence Feedback for Worker Productivity

Haoyuan Liu
Nanyang Technological University
haoyuan.liu@ntu.edu.sg

Wen Wen
University of Texas at Austin
wen.wen@mccombs.utexas.edu

Ashish Agarwal
University of Texas at Austin
ashish.agarwal@mccombs.utexas.edu

Andrew Whinston
University of Texas at Austin
abwhins@gmail.com

Abstract

With the rapid development of artificial intelligence (AI) technologies, many organizations have adopted AI to collect data on worker behavior and provide feedback to workers based on such data (for simplicity, we call such tools as AI supervisors). In this study we explore how workers' productivity is shaped by AI supervisors. We design and implement a large-scale randomized field experiment to quantify the economic impact of an AI supervisor on sales workers' productivity and distinguish its effect on work effectiveness vs. work efficiency. Our results show that the AI supervisor positively influenced bottom-ranked sales workers' productivity but had a negative impact on top-ranked workers' productivity. We further seek to understand the mechanisms through which AI feedback influenced sales workers: Bottom-ranked workers' productivity gain was driven by improvement in both selling effectiveness and customer engagement efficiency, whereas top-ranked workers' productivity loss was largely driven by their reduction in customer engagement efficiency.

Keywords: Artificial Intelligence (AI), AI Feedback, AI Supervisor, Worker Productivity, Randomized Field Experiment

1. Introduction

The use of artificial intelligence (AI) within organizations has recently become prevalent in a range of industries and functional areas. It has been playing an important role that traditional information technology (IT) systems have never been able to offer before – to manage and interact with workers instead of serving as merely a tool for workers. Such AI-human connections can come in different forms: from tracking the digital footprint of the workers for performance evaluation (e.g., the solutions offered by Controlio) to using webcams and computer vision technologies to monitor workers (e.g., the solutions offered by Drishti), from listening to call handler's response and giving real-time suggestion (e.g., the solutions offered by Cogito) to a fully automated fulfillment center (e.g., Amazon fulfillment centers) where human workers' job is to carry out AI's instructions. In the area of management and human resources, through collecting and analyzing a vast amount of data on worker behavior and performance, AI could help organizations to predict worker turnover, provide personalized training, and evaluate workers' job performance.

A growing body of literature studying the use of AI in managing workers suggests that AI can help improve worker productivity by generating evaluations or recommendations that are consistent, accurate, and systematic (e.g., Tong et al., 2021). However, there is a widespread fear that AI in management will lead to burnout, stress, and mental health issues, and deprive workers of their character (e.g., Roscigno & Hodson, 2004; Bernstein, 2012; Cater & Heikkilä, 2021).

In this paper, we focus on the type of AI tools used to closely track workers' daily behavior and provide them feedback based on the observed data (for simplicity, we call such tools as AI supervisors hereafter). We seek to understand how an AI supervisor can influence worker productivity via giving workers detailed and structured feedback. Moreover, motivated by a body of literature that looks at differential impact of AI based on worker characteristics (e.g., Luo et al. 2021), we seek to examine whether and how workers with different historical performance react to feedback given by their AI supervisor (hereafter denoted as AI feedback) differently.

To identify the causal effect of AI feedback on workers' productivity, we work with a drug store company in China to conduct a large-scale randomized field experiment. The company has a large set of chain stores across different cities and regions, and these stores sell a similar set of products including over-the-counter (OTC) drugs, dietary and nutrition supplements, and skincare products. We work with the company to evaluate the economic impact of an AI system that is used to provide feedback to its sales

URI: https://hdl.handle.net/10125/106606
978-0-9981331-7-1

workers (i.e., an AI supervisor in our term). More specifically, to interact with the AI supervisor, each sales worker will wear a badge with a built-in microphone, which records her conservations with customers. The recordings will then be uploaded to a cloud server where proprietary AI algorithms are used to analyze whether the conversations include six aspects of talking points that the management team has defined as important metrics for successful selling. Then, the feedback that indicates how frequently the worker has addressed each of the six aspects (among all customers she has talked to) will be sent back to the worker and her manager.

The randomized field experiment spanned six months in total. Among 90 stores that participated in our experiment, we randomly assigned half into the treatment group and the other half into the control group. For the first two months, none of the chain stores owned by the company had implemented the AI system. To disentangle the effect of AI feedback from the effect of AI monitoring, we divided the post-treatment period (a total of four months) for the treatment group into two phases. In the first phase which spanned two months, workers were required to wear the badge but not provided with feedback; therefore, in this phase workers would largely perceive themselves as being monitored. In the second phase, which also spanned two months, workers were not only required to wear the badge but also provided with feedback about their performance on the six aspects mentioned earlier. Based on the difference-in-differences empirical approach, we use the first phase of the post-treatment period to quantify the impact of AI monitoring and use the second phase to quantify the combined effect of AI monitoring and AI feedback. Then, the net effect of AI feedback could be inferred from subtracting the monitoring effect from the combined effect.

Our results can be summarized as follows. Overall, when we pool all workers together, there is mixed evidence on how the AI supervisor influenced workers. However, after we divide workers into different categories based on their past performance (i.e., top-ranked vs. bottom-ranked workers), the empirical evidence reveals very interesting and contrasting patterns on how the AI supervisor influenced different workers differently. On the one hand, the AI supervisors positively influenced bottom-ranked workers' productivity by improving both their weekly number of transactions and revenue per transaction. On the other hand, the AI supervisor negatively affected top-ranked workers' productivity, particularly regarding the number of transactions made each week.

To further understand the mechanisms through which the feedback provided by the AI supervisor influenced workers, we conduct a range of empirical analyses based on detailed behavioral data collected by the AI system during the post-treatment period for workers in the treatment group. First, we find that workers did react to the feedback by increasing the mention rates of the six aspects emphasized in the feedback report. However, regarding the aspects that could be important to selling but not covered in the feedback, bottom-ranked workers reduced the mention rate of these aspects whereas top-ranked workers kept the same level of mention rate as before. Second, for top-ranked workers, AI feedback did not improve their selling effectiveness in terms of both conversion rate and revenue per transaction, and at the same time, AI feedback dampened their selling efficiency, as reflected by a reduction in number of customers per hour they were able to engage. In contrast, for bottom-ranked workers, AI feedback seemed to significantly boost their selling effectiveness, as reflected by significant increases in both conversion rate and revenue per transaction; AI feedback also helped them to become more efficient, as they experienced an increase in number of customers engaged per hour.

2. Related Literature

This paper contributes to a growing body of literature that focuses on the use of AI in managing workers (e.g., Luo et al., 2021; Tong et al., 2021; Lou & Wu, 2021). Most relevant to our study are Luo et al. (2021) and Tong et al. (2021). In particular, Tong et al. (2021) studied the effect of AI feedback on employee performance. They argued AI increases employee productivity by increasing the quality of feedback when compared with human feedback; at the same time, AI also harms productivity, as workers may lack trust in AI once they know the feedback is provided by AI. Luo et al. (2021) explored AI coaches that provide training to sales agents and showed an inverted U-shape on the relationship between worker performance and their reaction to AI coaches. Similar to our research, they examined the heterogeneity in AI's effects based on the performance of sales agents. However, they focused on psychological factors including information overload and aversion that could influence workers' reaction to AI coach. In contrast to their finding on top workers' lack of trust in AI, we show that top-ranked workers did respond to AI feedback in our setting. In fact, top-ranked workers not only sought to improve on the aspects stressed by the feedback but also tried to keep their own preferred selling styles and tactics, most of which were not covered by the report. As a result of such a higher workload for each conversation with the customers, top workers ended up engaging with fewer customers

per hour. In sum, while both Luo et al. (2021) and our study showed top-ranked workers seemed to be hurt by AI, our work complements their study by highlighting different underlying mechanisms through which AI feedback shapes worker behavior.

More broadly, our research is related to the stream of literature that seeks to understand the relationship between IT and productivity (e.g., Aron et al., 2011; Tambe & Hitt, 2012). By leveraging the detailed worker-level behavioral data collected by the AI system, we are able to identify the effectiveness of AI feedback in changing worker behavior and uncover detailed mechanisms through which workers respond to AI feedback. Consistent with the existing literature on the standardization effect of IT in business processes, we observe a similar effect of AI feedback. Particularly for bottom-ranked workers, they increased their efforts on addressing the specific six aspects outlined in the feedback but reduced efforts on other aspects that are related to the selling process but not covered in the feedback. As a result, the conversations carried out with customers could become more standardized.

3. Empirical Approach

3.1. Empirical setting

We work with a national drug store retailer to evaluate the effects of AI supervisors on sales workers' productivity. The company has chain drug stores distributed across different cities and regions. All its chain stores are owned by the same company and thus have the same business processes; they all sell the same wide selection of products such as over-the-counter (OTC) drugs, dietary and nutrition supplements, and skincare products. Typically, each store has around five employees, all of whom are qualified to sell every item within the store. Other than a point of sale (POS) system that records the number of transactions and the associated sales revenue made by each sales worker, these stores do not have any other devices to track sales worker's behavior or performance.

The AI system considered by the company includes both physical hardware and software. The hardware is a badge that a sales worker would wear during their working time. The badge has an array of built-in microphones, used to record the sales worker's conversations with customers; it also provides enough storage space and battery to record all the conversations during the working hours of a day. At the end of each day, a worker would charge the badge at a charging station that also automatically uploads the recorded data to a cloud server, where the data are analyzed.

To protect customer privacy, customer voice data are deleted; only the voice data of sales workers are fed into some proprietary natural language processing (NLP) algorithms running in the cloud. For each conversation, the algorithms predict whether a given sales worker addressed each of the following six aspects and assign a dummy score of either one or zero (denoted as *six markers* hereafter). The six markers are: 1) showing empathy for customer's well-being, 2) being positive in fulfilling customer's needs, 3) proactively identifying customer issues and needs, 4) suggesting alternatives, 5) cross-selling, and 6) ensuring proper consumption of the purchased item(s). They are defined by the drug store retailer based on its understanding of the market and important factors leading to successful sales in the past. This NLP model had been trained and yielded satisfactory results in terms of voice recognition and detection of the six markers before our experimental period.

After the data are analyzed, a feedback report is generated for each sales worker, showing the frequency of addressing each of the six markers among all conversations during a given period. For example, suppose a worker sought to do cross-selling 50 times among the conversations with 100 customers during a certain period, then, the mention rate of this marker (i.e., cross-selling) is 0.5, and the worker would receive a report that shows such a mention rate for the cross-selling aspect. Due to the requirement by the management team of the drug store retailer, the report is first sent to a worker's store manager, who in turn forwards it to the worker. The content of the feedback report only includes the mention rate of each of the six markers individually and the average mention rate across all six markers, a total of seven metrics.

In addition to the six markers, the algorithms also capture the mention rates of other aspects relevant to the selling process but not included in the reports sent to the stores. These unreported markers include introducing best-seller products, introducing sales campaigns, introducing bundled goods, comparing with competing products, providing price matching, encouraging customers to revisit, helping customers to register for membership, introducing procedures for chronicle diseases medicine refills, and helping customers redeem coupons and gift cards. However, because the feedback does not cover the mention rates of these aspects, both store managers and workers would not know the fact these aspects are also captured by the AI supervisor system.

3.2. Field experiment and empirical strategy

To identify the causal impact of such an AI supervisor on sales workers' productivity, we worked

with the drug store retailer and designed the following randomized field experiment. A total of 90 stores with a total of 481 sales workers were included in the study. These stores had similar sizes (in terms of the number of sales workers) and sales revenue when the experiment took place. They came from two cities, with 38 stores in City 1 and 52 stores in City 2. In City 1, 19 stores were randomly selected into the treatment group for which the AI system would be implemented, and the rest 19 stores were used as the control group for which the AI system would not be adopted for the entire experimental period. Similarly, in City 2, we randomly selected 26 stores as the treatment and the remaining 26 stores as the control group. The stores across the treatment and the control groups were under the same managerial administration, and we were able to work with the company to ensure no other systematic changes would happen for the treatment group during the experimental period except the AI supervisor system implementation.

Existing literature suggests the potential benefits from IT-based monitoring systems such as increasing fairness and reducing mild forms of misconduct such as shirking and absenteeism (e.g., Hubbard 2000; Baker and Hubbard 2003; Duflo et al. 2012; Pierce et al. 2015; Staats et al. 2017). Then, in the case of AI, given that it both monitors workers to collect worker data and provides feedback to workers, the observed changes in productivity after the deployment of AI could be due to a combination of its monitoring role and feedback role. To identify the latter, which is our main focus of this study, we worked with the company to implement the experiment with multiple phases. In particular, our experimental period started on October 1, 2020 and ended on March 31, 2021, a total of six months. From December 1st, 2020, to January 31st, 2021, the treatment group went through the first phase of the post-treatment period (hereafter denoted by P1). During this phase, sales workers in the treatment stores were required to wear the badge and charge them at the charging station after work every day. However, during this phase, no feedback reports were generated. Although we did not specifically tell the stores the purpose of such a system, they largely assumed it was used to monitor their behavior. From January 1st, 2021 to January 31st, 2021, the system was fully operational in the treatment stores, so we were able to obtain all data about these sales workers' behavior, though such data were not shared with the stores.

The second phase of the treatment started on Monday February 1st, 2021, when the first set of reports were sent to each store in the treatment group (hereafter denoted by P2). As mentioned above, the report included the mention rate of each of the six markers and the average mention rate across all six markers. This first batch of reports was generated based on workers' conversation data with customers in January 2021. On February 15th, the second batch of reports regarding workers' behavior from February 1st to February 14th were generated and distributed. After February 15th, new reports were generated and shared on a weekly basis, each covering workers' behavior in the previous week.

Based on this experimental design, we could largely assume workers in the treatment group would feel being monitored during the first phase and once they started to receive feedback on February 1st, 2021 (the first day of the second phase), they would be affected by both being monitored by the system and AI's feedback (i.e., the specific suggestions provided by the system).

To identify the causal impact of AI feedback on a worker's productivity, we use the following difference-in-differences model, where our unit of analysis is at the sales worker (denoted as i) – week (denoted as t) level:

$$Productivity_{it} = \beta_1 \times Treated_i \times P1_t + \beta_2 \times Treated_i \times P2_t + \eta_t + v_i + \varepsilon_{it} \quad (1)$$

$Productivity_{it}$ denotes sales worker i's productivity in week t. We use two metrics to measure a worker's productivity: 1) the number of transactions made by the worker in week t, and 2) average revenue per transaction in week t. The former is determined by both the level of customer engagement (i.e., how many customers a worker carries out conversations, an efficiency measure) and the conversion rate (which directly relates to a worker's selling skills, an effectiveness measure). The latter is mostly influenced by a worker's selling effectiveness, particularly if she could successfully do cross-selling or upselling.

The dummy variable $Treated_i$ is equal to one if a worker i is in the treatment group and zero if the worker is in the control group. The time dummy variable $P1_t$ is turned on during first phase when the AI system only played a monitoring role. That is, it is equal to one from December 1st, 2020, to January 31st, 2021, and equal to zero otherwise. $P2_t$ is equal to one from February 1st, 2021 to March 31st, 2021, and equal to zero otherwise. η_t is a set of weekly dummies to control for the general time trend across the treatment group and control group. Due to the inclusion of η_t, the direct effects of $P1_t$ and $P2_t$ would not be estimated. To control for time-invariant worker-level characteristics, we include worker-fixed effects for all specifications (denoted as v_i).

The coefficient β_1 captures the effect of AI monitoring on the treatment group during the first phase, whereas β_2 captures the combined effect of AI monitoring and AI feedback during the second phase. As a result, we could subtract β_1 from β_2 to identify the

effect of AI feedback. One important concern is whether the effect of AI monitoring changed from P1 to P2. We believe if there were a non-constant effect, it would be most likely to diminish over time due to the greater familiarity with the system. Therefore, this may lead to an underestimation of the positive effect of AI feedback.

To understand how workers with different historical performance react to an AI supervisor differently, we decompose workers into three categories—*top-ranked worker*$_i$, *mid-ranked worker*$_i$, and *bottom-ranked worker*$_i$—based on her past performance, measured by average weekly sales revenue prior to the treatment date for both the treatment group and the control group.

More specifically, after obtaining the distribution of pre-treatment-period average weekly sales revenue for all workers in our sample, workers with sales revenue in the top quartile of the distribution are classified as top-ranked workers (i.e. *top-ranked worker*$_i$ would be equal to one); workers with sales revenue in the bottom quartile of the distribution are classified as bottom-ranked workers (i.e. *bottom-ranked worker*$_i$ would be equal to one); the rest of the workers are then considered as mid-ranked workers (i.e. *mid-ranked worker*$_i$ would be equal to one). Then, we interact these three dummy variables with both *P1* and *P2* to control for the overall time trend on changes in productivity for workers in the treatment and control groups in different performance bracket. Our key variables of interest are the interactions between these performance bracket dummies and *P1*Treated* or *P2*Treated.* The coefficient estimates of these three-way interaction terms are used to identify how workers in the treatment group with different performance responded to AI monitoring and AI feedback differently.

$$Productivity_{it} = \sum_{j=1}^{3} \beta_{1j} \times Treated_i \times Performance\ bracket_{ij} \times P1_t + \sum_{j=1}^{3} \beta_{2j} \times Treated_i \times Performance\ bracket_{ij} \times P2_t + \sum_{j=1}^{3} \beta_{3j} \times Performance\ bracket_{ij} \times P1_t + \sum_{j=1}^{3} \beta_{4j} \times Performance\ bracket_{ij} \times P2_t + \eta_t + v_i + \varepsilon_{it} \quad (2)$$

4. Results

4.1. Comparison between treatment group and control group during pre-treatment period

To check the comparability between the treatment group and the control group, we compare workers in the treatment group against workers in the control group on key productivity metrics (i.e., number of transactions and revenue per transaction) and demographics (i.e., age and tenure) prior to the treatment. As shown in Table 1, a worker in the treatment group made 124 transactions each week, and each transaction was worth 55 RMB; similarly, a worker in the control group had 128 transactions each week and each transaction worth 53 RMB. There is no statistically significant difference in both variables between the treatment and the control group. Workers from the treatment and control group also have similar ages and tenure (i.e., years spent with the company).

Table 1. Comparison between treatment group and control group, pre-treatment period

Variables	Obs.	Treatment group	Control group	Difference	p-value
Weekly revenue	3,169	8139.111 (171.699)	8374.396 (179.465)	-235.285 (248.342)	0.344
Revenue per transaction	3,169	54.894 (1.218)	53.136 (0.947)	1.758 (1.544)	0.255
Weekly No. of transactions	3,169	124.266 (2.483)	128.053 (2.447)	-3.786 (3.486)	0.278

Notes: The variables are measured at the individual-week level.

In addition, we perform a parallel assumption check. Averaging the two key dependent variables - revenue per transaction and the number of transactions - at a weekly level, we present the raw trend of dependent variables of the treatment and the control group in Figure 1. As can be seen in the figure, prior to the treatment (hereby called P0), which is shown to the left of the first vertical line, the treatment and the control group show a very similar trend. After the first phase of the treatment started, the number of transactions showed some difference. Such difference becomes more noticeable after the second phase of the treatment when the AI feedback was being provided.

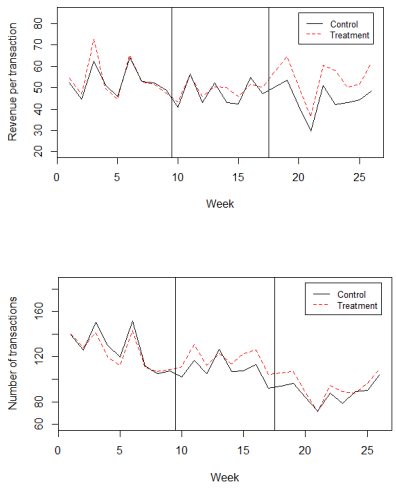

Figure 1. Parallel assumption of the sample

4.2. Baseline results

The results based on specification (1) are shown in Table 2, where we decompose productivity by revenue per transaction and the number of transactions. While revenue per transaction may indicate the selling effectiveness, or selling skills, the number of transactions is a combined result of worker efficiency and selling skills (i.e., that relates to conversion). As suggested by the last row of Table 2, overall, there was no significant increase in the number of transactions caused by AI feedback, but we do observe a significant increase in revenue per transaction as a result of the AI feedback.

Table 2. Baseline results

Dependent variable	No. of transactions	Revenue per transaction
	(1)	(2)
Treated*P1	8.208*	0.604
	(4.443)	(1.921)
Treated*P2	3.610	4.694*
	(5.620)	(2.582)
Number of observations	10,772	10,772
Number of sales workers	481	481
Adjusted R-squared	0.062	0.037
Marginal effect of AI feedback	-4.598	4.090**
	(4.365)	(2.114)

Notes: Robust standard errors clustered at the worker level are in parentheses. All regressions include worker-fixed effects and time-fixed effects (i.e., weekly dummies). *** p < 0.01, ** p < 0.05, * p < 0.1.

Table 3 presents the heterogeneous effects of AI feedback based on workers' historical performance, where there are very interesting and contrasting patterns among top-ranked workers vs. bottom-ranked workers. As shown in the last few rows of Table 3 regarding the marginal effect of AI feedback on top-ranked workers, top-ranked workers seemed to suffer productivity loss after they were provided with AI feedback, as reflected by a significant decrease in the number of transactions. One plausible explanation is that such AI feedback imposes some standardized evaluation metrics for all workers to follow. It may not only prevent these top-ranked workers from fully utilizing their capabilities (e.g., Oliver & Anderson 1994; Ahearne et al., 2010; Boone & Özcan, 2014) but also potentially lead to some work inefficiency. As we will discuss in greater detail in the next few sections, we implement a range of analyses to identify potential mechanisms that drive such a decrease in number of transactions made by top-ranked workers after the introduction of AI feedback.

Meanwhile, as indicated in column (2) of Table 3, there was no significant change in revenue per transaction generated by top-ranked workers after they were given the feedback. As noted earlier, revenue per transaction is mostly determined by a salesperson's selling effectiveness, i.e., selling skill. A plausible explanation of this result is that because AI feedback was partly derived from the past best practice of most successful workers, it may not be very effective on top-ranked workers since they have already done well on these metrics (e.g., MacLean & Behnam, 2010).

In contrast, an investigation of bottom-ranked workers' productivity changes reveals the opposite pattern. As shown in the last row of Table 3, after receiving AI feedback, bottom-ranked workers had an improvement in the number of transactions. Moreover, AI feedback also boosted their productivity regarding selling effectiveness, as reflected by a significant increase in revenue per transaction. A plausible explanation is that these bottom-ranked workers did need constant reminder on how to improve their selling skills (e.g., Oliver & Anderson 1994). As a result, they benefit from AI feedback significantly.

Table 3. Heterogeneous effect on workers with different historical performance

Dependent variable	No. of transactions	Revenue per transaction
	(1)	(2)
P1*Treated * Top-ranked worker	-3.028	-1.813
	(10.379)	(2.858)
P2*Treated * Top-ranked worker	-23.842*	-1.117
	(13.666)	(4.803)
P1*Treated * Mid-ranked worker	11.857**	0.666
	(5.752)	(2.247)
P2*Treated * Mid-ranked worker	5.701	2.864
	(6.881)	(3.412)
P1*Treated * Bottom-ranked worker	12.024*	2.792
	(6.945)	(5.272)
P2*Treated * Bottom-ranked worker	23.155***	13.136**
	(7.580)	(5.778)
Number of Observations	10,772	10,772
Number of sales workers	481	481
Adjusted R-squared	0.087	0.040
Marginal effect of AI feedback for top-ranked workers	-20.814*	0.696
	(11.330)	(4.776)
Marginal effect of AI feedback for bottom-ranked workers	11.130*	10.344***
	(6.414)	(3.642)

Notes: Robust standard errors clustered at the worker level are in parentheses. All regressions include worker-fixed effects and time-fixed effects (i.e., weekly dummies). All regressions include the two-way interactions among P1/P2 and Top/Mid/Bottom-ranked workers. Due to the limited space, the coefficient estimates of these two-way interaction terms are not reported in the table but are available upon request. *** p < 0.01, ** p < 0.05, * p < 0.1.

4.3. How AI feedback influences workers' productivity

In the previous section, we show that top-ranked workers and bottom-ranked workers reacted to AI feedback very differently. In this section, we further break down productivity into more detailed measures. This breakdown, distinguishing between effectiveness and efficiency at selling, could shed more light on how different workers learn and react to AI feedback. To do so, we leverage the detailed behavioral data the AI supervisor collected from the workers. In particular, the total productivity of a sales worker can be broken down into revenue per transaction, the conversion rate, and the hourly number of customers engaged. While revenue per transaction and conversion rate reflects

the effectiveness of a sales worker, the hourly number of customers engaged could reflect the efficiency of a sales worker.

One limitation of the data is that because AI supervisor was installed during phase 1 of the experiment only for the treatment group but not for the control group, we are unable to implement a difference-in-differences strategy. However, because each store in the treatment group is matched with one store in the control group from the same city's nearby area, we use the number of transactions and revenue per transaction of the matched store in the control group to control for demand seasonality and spending seasonality faced by a focal store in the treatment group. Meanwhile, because we do not have worker's behavioral data before phase 1, our analyses in this section and the following section focus on comparing worker's behavior between phase 1 and phase 2. We believe it is reasonable to assume that the effect of monitoring on selling efficiency and effectiveness remains the same between the two phases, as merely monitoring without any suggestions or feedback can hardly influence selling tactics. As a result, the changes in selling efficiency and effectiveness from phase 1 to phase 2 would be mostly attributable to the provision of AI feedback.

The results on the impact of AI feedback on efficiency and effectiveness are presented in Table 4. As shown in the first two columns where we use the number of customers engaged per hour as a proxy for selling efficiency, although there was no significant change overall (as reflected in column [1]), there is important heterogeneity across workers with different ranks, as shown in column (2). In particular, bottom workers experienced an increase in the hourly number of customers engaged, whereas top workers had a significant decrease in hourly number of customers engaged. One plausible explanation is that for bottom workers, with AI feedback that specifically highlights the six markers important to selling, they have much clearer goal in terms of what should be mentioned during the selling process, thereby being more efficient in dealing with customers. On the other hand, because top workers usually have higher self-efficacy, they may be less willing to change their work style and habits (e.g., Tarakci et al., 2018). As a result, while they may work on the six markers highlighted in the report, they may also keep their other selling tactics that are not in the report but are believed to be useful due to their success in the past. This multitasking could lead to reduced efficiency. Another potential explanation for such downward efficiency could be that AI feedback is likely to cause top workers, who are more strategic, to shift their attention from engaging more customers to engaging fewer customers but keep every engagement high quality and address the six reported markers to the best of their ability. In other words, since customer engagement efficiency was not emphasized by the AI feedback, top workers may trade efficiency for higher conversion and revenue per transaction, as well as higher marker mention rate, for strategic consideration. In the next section, we will implement some additional analyses to investigate whether these explanations hold.

Columns (3) through (6) of Table 4 show how AI feedback influenced conversion rate and revenue per transaction, two metrics used to capture selling effectiveness. The results suggest that AI feedback had a particularly strong effect on the selling effectiveness of bottom-ranked workers whereas there was little influence on top-ranked workers. This is probably because AI feedback, due to its structured, systematic, and thus actionable nature, may help bottom-ranked workers to learn selling skills quickly without trial and error (e.g., Anderson, 1987; Keith & Frese, 2008). However, there is little room for top-ranked workers to further improve these metrics, as they may have already done well.

Table 4. How does AI feedback affect sales workers' efficiency vs. effectiveness?

Dependent variable	Hourly no. of customers engaged		Revenue per transaction		Conversion rate	
	(1)	(2)	(3)	(4)	(5)	(6)
P2	-0.125		6.679**		0.033	
	(0.097)		(3.305)		(0.021)	
P2*Top worker		-0.416***		3.442		-0.009
		(0.129)		(4.921)		(0.039)
P2*Mid worker		-0.199		5.771		-0.005
		(0.125)		(3.730)		(0.025)
P2*Bottom worker		0.357*		12.313**		0.183***
		(0.195)		(5.184)		(0.047)
Demand seasonality	52.230***	51.732***	3.823	3.728	0.136**	0.133**
	(12.938)	(12.943)	(11.191)	(11.210)	(0.060)	(0.060)
Spending seasonality	47.349***	48.809***	97.458***	97.748***	0.320***	0.330***
	(17.434)	(17.490)	(12.371)	(12.440)	(0.100)	(0.100)
Adjusted R-squared	0.079	0.093	0.048	0.051	0.013	0.031

Notes: Robust standard errors clustered at the worker level are in parentheses. All regressions include worker-fixed effects. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Number of observations: 2093; number of sales workers: 188.

4.4. Potential mechanism through which AI feedback influences worker efficiency

As noted above, the differential effect of AI feedback on work efficiency, as measured by the hourly number of customers engaged, is particularly counter-intuitive. While bottom workers gained efficiency, as what we would hope to get from the AI supervisor, top workers lost their efficiency after receiving AI feedback. Such a decline in efficiency may be due to top workers' strategic shift of attention, or due to multitasking. In this section, we will provide some suggestive evidence to examine whether these mechanisms hold.

One explanation for the downward efficiency of top workers could be that these top workers are strategic. Since the report focuses on the six markers' mention rates but does not include measures for customer engagement efficiency such as hourly

number of customers engaged, these workers may intentionally reduce the interactions with potential customers (i.e., to reduce the number of conversations, which is used as the denominator to calculate the mention rate) but make sure to address as many markers as possible for each interaction. Such reduced customer engagement could lead to a loss in productivity in terms of the number of transactions.

However, a closer look at the results shown in Table 3 and Table 4 suggests that this explanation may not hold. More specifically, if these top workers behaved strategically, while they might reduce the level of customer engagement in order to achieve higher mention rates shown in the report, they would improve their conversion rate for each customer they engaged so that they can achieve similar number of transactions, as the number of transactions is one of key performance metrics used by the company for evaluation and promotion purposes. However, Table 4 shows their conversation rate did not increase from phase 1 to phase 2. Based on column (1) of Table 3, top-ranked workers in fact faced a decline in the number of transactions, because of a decline in customer engagement and unchanged conversation rate. In addition, the top-ranked workers did not have an increase in revenue per transaction, suggesting it was unlikely that they strategically shifted their focus to more profitable customers. Overall, the worsened performance of top workers seemed contradictory to the explanation that they were behaving strategically.

As noted earlier, another possible explanation for such reduced customer engagement by top-ranked workers is multitasking. These top workers might not only want to improve the mention rates of the six reported markers but at the same time keep their preferred selling style and tactics, most of which would not be covered by the report. As a result, due to a higher workload for each conversation with the customers, top workers ended up engaging with fewer customers per hour.

To test whether this explanation holds, we first examine whether top workers did multitask after receiving AI feedback, i.e., whether they both increased mention rates of reported markers but also kept the same level of other markers not reported. The results are shown in Table 5.

As shown in columns (1) and (3), we find that overall, the mention rate of six markers reported in the feedback significantly increased but that of the unreported markers significantly decreased. As shown in columns (2) and (4) when we look at how mention rates changed based on different performance brackets, we find that, regardless of performance brackets, all workers increased their mention rate of the six reported markers. This suggests that workers did react to AI feedback strongly regardless of the historic performance of the worker. However, bottom workers reduced the mention rate of unreported markers significantly, but top workers kept the same level of mention rate of these unreported markers. This piece of evidence seems to be consistent with the explanation that top workers could experience a heavier workload per customer conversation.

Table 5. Do workers multitask after receiving AI feedback?

Dependent variable:	Six markers mention rate		Unreported markers mention rate	
	(1)	(2)	(3)	(4)
P2	0.025***		-0.012***	
	(0.009)		(0.004)	
P2*Top worker		0.010*		0.008
		(0.006)		(0.007)
P2*Mid worker		0.021**		-0.010**
		(0.009)		(0.004)
P2*Bottom worker		0.051**		-0.034***
		(0.025)		(0.008)
Demand seasonality	0.006	0.006	-0.017*	-0.017*
	(0.020)	(0.020)	(0.009)	(0.009)
Spending seasonality	-0.014	-0.013	0.039***	0.038***
	(0.020)	(0.020)	(0.011)	(0.011)
Adjusted R-squared	0.024	0.031	0.021	0.055

Notes: Robust standard errors clustered at the worker level are in parentheses. All regressions include worker-fixed effects. *** p < 0.01, ** p < 0.05, * p < 0.1. Number of observations: 2093; number of sales workers: 188.

After establishing some evidence on top workers' multitasking behavior, we next seek to understand whether multitasking would indeed lead to efficiency loss. To do so, we focus on the subsample of top-ranked workers. We identify the set of top-ranked workers who had increased the overall mention rate of all reported and unreported markers. Those workers would be the ones who engaged with multitasking the most. Then, we seek to identify whether those workers experienced the most reduction in the hourly number of customers engaged.

The results in Table 6 show that workers with a greater increase in overall mention rate of all reported and unreported markers did experience the most loss in efficiency.

Table 6. Does multitasking lead to efficiency loss for top workers?

Subsample of top-ranked workers	
Dependent variable: Hourly no. of customers engaged	
P2	-0.124
	(0.133)
P2* Increase in mention rate across all reported and unreported markers	-0.712**
	(0.267)
Demand seasonality	2.832***
	(0.418)
Spending seasonality	-1.006
	(0.677)
Number of Observations	420
Number of sales workers	40
Adjusted R-squared	0.221

Notes: Robust standard errors clustered at the worker level are in parentheses. All regressions include worker-fixed effects. *** p < 0.01, ** p < 0.05, * p < 0.1.

Overall, our analyses in this section seem to confirm the explanation that top workers' reduction in

customer engagement could be driven by their tendency to both address the feedback-related markers and preserve their preferred selling tactics.

5. Conclusion

In this study, we seek to understand how AI feedback influences worker's productivity. Based on a large-scale randomized field experiment, we discover important heterogeneity regarding how workers with different past performance reacted to AI feedback differently. Bottom-ranked workers benefited from AI feedback significantly. They were able to not only improve work efficiency, as measured by the hourly number of customers engaged, but also improve selling effectiveness, as captured by conversation rate and revenue per transaction. This seems to suggest for those workers, AI feedback does offer some structured and systematic approach for them to quickly learn skills without trial and error in a relatively short period of time (e.g., Anderson, 1987; Keith & Frese, 2008). In the meanwhile, bottom workers tend to have lower self-efficacy because of their relative performance position in the organization. As a result, they are likely to discard their own existing (perhaps unsuccessful) tactics (e.g., Tarakci et al., 2018) and are more willing to learn the best practices, especially when it is provided in great detail (e.g., Song et al., 2018). In other words, AI feedback may help these workers shift their focus to important aspects needed for successful selling and avoid unnecessary conversations with customers, thereby improving their efficiency.

On the other hand, top-ranked workers suffered a productivity loss from such AI feedback significantly. Such a loss in productivity was particularly driven by a reduction in work efficiency, as measured by the hourly number of customers engaged. Based on a set of analyses, we speculate the multitasking behavior they exhibited could be the underlying driving force—these workers not only wanted to follow AI feedback but also tended to keep their own unique skill sets and strategies (e.g., Dugosh & Paulus, 2005; Harrison & Rouse, 2015; Tarakci et al., 2018; North, 2019). As a result, they had to spend more time with each customer. Surprisingly, more time spent with each customer did not translate into a higher conversion rate and higher revenue per transaction for those top workers.

Our results have important implications for organizations that consider adopting AI supervisor systems. Our study highlights the need for managers to consider the heterogeneous reactions from different performance brackets of their workforce. For bottom-ranked performers or newer workers, AI feedback may serve as a helpful tool to train them. Yet for the top-ranked workers who may have heterogenous tactics that lead to success, managers should consider different approaches to mitigate the dampening effect of AI supervisor on their productivity. For example, managers may want to suggest to the top-ranked workers that AI supervisors are not to regulate their behavior, and that following the AI feedback or not would not be part of the evaluation process of the workers' performance.

6. References

Ahearne, M., Lam, S.K., Mathieu, J.E. and Bolander, W., 2010. Why are some salespeople better at adapting to organizational change? Journal of Marketing, 74(3), pp.65-79.

Anderson, J.R., 1987. Skill acquisition: Compilation of weak-method problem situations. Psychological review, 94(2), p.192.

Aron, R., Dutta, S., Janakiraman, R., Pathak, P.A., 2011. The Impact of Automation of Systems on Medical Errors: Evidence from Field Research. Information Systems Research 22(3):429-446.

Baker, GP, Hubbard, TN, 2003. Make versus buy in trucking: Asset ownership, job design, and information. American Economic Review 93(3):551–572.

Bernstein, E.S., 2012. The transparency paradox: A role for privacy in organizational learning and operational control. Administrative Science Quarterly, 57(2), pp.181-216.

Boone, C. and Özcan, S., 2014. Why do cooperatives emerge in a world dominated by corporations? The diffusion of cooperatives in the US bio-ethanol industry, 1978–2013. Academy of Management Journal, 57(4), pp.990-1012.

Cater, L., and Heikkilä, M., 2021. Your boss is watching: How AI-powered surveillance rules the workplace.

Cannon, M.D. and Witherspoon, R., 2005. Actionable feedback: Unlocking the power of learning and performance improvement. Academy of Management Perspectives, 19(2), pp.120-134.

Duflo E, Hanna R, Ryan S, 2012. Incentives work: Getting teachers to come to school. American Economic Review. 102(4):1241–1278.

Dugosh, K. L., & Paulus, P. B. 2005. Cognitive and social comparison processes in brainstorming. Journal of Experimental Social Psychology, 41: 313–320.

Goodman, J. S., Wood, R. E., & Chen, Z. 2011. Feedback specificity, information processing, and transfer of training. Organizational Behavior and Human Decision Processes, 115: 253–267.

Haas, M.R. and Hansen, M.T., 2005. When using knowledge can hurt performance: The value of organizational capabilities in a management consulting company. Strategic management journal, 26(1), pp.1-24.

Harrison, S.H. and Rouse, E.D., 2015. An inductive study of feedback interactions over the course of creative projects. Academy of Management Journal, 58(2), pp.375-404.

Keith, N., & Frese, M. 2008. Effectiveness of error management training: A meta-analysis. Journal of Applied Psychology, 93: 59–69

Kluger, A.N. and DeNisi, A., 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological bulletin, 119(2), p.254.

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

MacLean, T.L. and Behnam, M., 2010. The dangers of decoupling: The relationship between compliance programs, legitimacy perceptions, and institutionalized misconduct. Academy of Management Journal, 53(6), pp.1499-1520.

North, M.S., 2019. A GATE to understanding "older" workers: Generation, age, tenure, experience. Academy of Management Annals, 13(2), pp.414-443.

Oliver, R.L. and Anderson, E., 1994. An empirical test of the consequences of behavior-and outcome-based sales control systems. Journal of marketing, 58(4), pp.53-67.

Parker, S.K., Ward, M.K. and Fisher, G.G., 2021. Can high-quality jobs help workers learn new tricks? A multidisciplinary review of work design for cognition. Academy of Management Annals, 15(2), pp.406-454.

Pierce, L., Snow, D.C. and McAfee, A., 2015. Cleaning house: The impact of information technology monitoring on employee theft and productivity. Management Science, 61(10), pp.2299-2319.Roscigno, V.J., Hodson, R., 2004. The organizational and social foundations of worker resistance. American Sociology Review 69(1):14–39.

Song, H., Tucker, A.L., Murrell, K.L. and Vinson, D.R., 2018. Closing the productivity gap: Improving worker productivity through public relative performance feedback and validation of best practices. Management Science, 64(6), pp.2628-2649.

Staats, B.R., Dai, H., Hofmann, D. and Milkman, K.L., 2017. Motivating process compliance through individual electronic monitoring: An empirical examination of hand hygiene in healthcare. Management Science, 63(5), pp.1563-1585.

Tambe P., Hitt LM., 2012. The productivity of information technology investments: New evidence from IT labor data. Information Systems Research, 23(3-part-1), pp.599-617.

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