AI, Skill, and Productivity: The Case of Taxi Drivers*

Kyogo Kanazawa[§] Daiji Kawaguchi[§†]
Hitoshi Shigeoka[‡] Yasutora Watanabe[§]

[§]University of Tokyo
[†]RIETI and IZA
[‡]Simon Fraser University and NBER

May 15, 2021

Preliminary and Incomplete. Comments are more than welcome.

Abstract

Artificial Intelligence (AI) is documented to have differential impacts *across* occupations. However, micro-level evidence is scant on how AI affects the productivity of workers with different skill levels *within* occupation. We study the impact of a demand-forecasting AI on productivity in the context of taxi drivers. We find that the AI improves drivers' productivity by shortening the time to search for customers by 5% on average. Importantly, the productivity gain is concentrated on low-skilled drivers (10%) whereas the corresponding gain on high-skilled drivers is nearly zero. This result suggests that AI, unlike past technologies, may not accompany skill-biased technological change.

Keywords: Artificial Intelligence, Skill, Productivity, Taxi-drivers, Prediction, Demand forecasting, Machine learning, Skill-biased technological change

JEL codes: J22, J24, L92, R41

*We thank Kensuke Teshima, and seminar participants at Hitotsubashi University, Nihon University, Osaka University for their excellent comments and suggestions. The data used in this paper is provided by a tech company that developed the AI application but we have no financial supports from the company or any other conflicts of interest. This research is financially supported by Japan Science and Technology Agency (JPMJRX18H3). Kanazawa, kanazawa@e.u-tokyo.ac.jp; Kawaguchi, kawaguchi@e.u-tokyo.ac.jp; Shigeoka, hitoshi_shigeoka@sfu.ca; Watanabe, yasutora.watanabe@gmail.com

1 Introduction

Artificial Intelligence (AI) has a potential to drastically reshape employment (Brynjolfsson et al. 2018).[1] However, the impact of AI on employment could be fundamentally different from those of the past technologies including IT and robotics, which are considered to be skill augmenting and inequality enhancing (Autor et al. 2003; Bartel et al. 2007; Acemoglue and Restrepo 2020). Whereas past technologies have replaced manual and routine tasks, AI will replace non-routine cognitive tasks. In fact, Webb (2020) documents that AI technology mostly affects *high-skilled* occupations as an AI substitutes for tasks that require the type of skill that high-skilled workers possess.

All the studies examining the impact of AI consider differential impact *across* occupations based on the degree to which each occupation is exposed to AI (Felten et al. 2018, 2019; Frank et al. 2019; Webb 2020; Alekseeva et al. 2020). As a result, these studies implicitly assume that all the workers *within* the occupation are *uniformally* affected by the AI.[2] However, even within occupation, there is a substantial heterogeneity of skill for the task that can be replaced by AI. Thus, examining the impact of AI on workers with different skill within an occupation provides a deeper understanding of AI's complex impact on human labor at granular level. However, such evidence from micro-level data is scant.[3]

To fill in the gap, we study the impact of AI on productivity across different worker skills in the context of taxi drivers. Taxi drivers are an ideal case to answer this question for several reasons. First, worker's individual productivity is easily measured by the length of time to search for customers as well as sales. This is because each driver works independently and has considerable discretion as to how they search for customers. Our data show that more than 80% of driver's working time is devoted to searching for customers (and the remaining to carrying the customers, excluding breaks). Thus, searching for customers is among the most important task for taxi drivers, and hence the length of search time is an important productivity measure. Second and relatedly, since productivity is well-defined, the construction of appropriate metrics of skill is also straightforward; we can define the skill based on the past performance that impacts the same productivity measure (i.e., length of

[1] Frey and Osborne (2013, 2017) predict that AI may replace 47% of current jobs in the US in 10-20 years.

[2] These studies are based on a task-based model of technology and labor (Acemoglue and Restrepo 2018; Webb 2020), where each occupation consists of various tasks and automation occurs at the task level. In this framework, the occupation in which high proportion of tasks can be replaced by AI is considered to be highly exposed to AI. Each study differs in the way which tasks are considered to be replaced by AI.

[3] The only exception to our knowledge is Grennan and Michaely (2020), which study the impact of AI on security analysts. They document that analysts are more likely to leave the profession when they cover stocks that are more exposed to AI because data for such stocks are more abundantly available. They find more pronounced effects on accurate analysts.

search time). Such constructed skills precisely capture the skill level of workers unlike other proxies for a skill such as potential experience or education, which are widely used in past studies (e.g., Katz and Murphy 1992; Card and Lemieux 2001; Autor et al. 2008). Third, the work environment of taxi drivers offers a clean setting to study productivity because all drivers work in a very similar setting; the taxi drivers possess the same capital, and charge the same prices, in the absence of the other confounding factors of productivity gain like differential capital input, and input prices.[4] Finally, operating a taxi involves various tasks whose substitutability with AI are considerably different: ranging from driving a vehicle on a narrow and congested city street (a difficult task for AI) to predicting the location of a customer (an easy task for AI).

The particular AI we study is called "AI Navi", which helps drivers search for the customers when a taxi is empty. The AI suggests routes based on the predicted demand to maximize the probability a taxi will catch customers given the current location. Thus, this AI is expected to improve the productivity of drivers by reducing the time they search for customers. This type of AI, which increases the accuracy of *prediction* tasks using machine learning techniques, is widely used in the real business setting (Agrawal et al. 2018, 2019). Demand forecasting as in this AI—the process of making estimations about future customer demand—is one of many existing prediction tasks. To the extent that such demand-forecasting skill is an important component of taxi drivers' skill set, the impact of AI may differ by the skill of drivers. In this paper, we test whether this AI improves worker productivity across drivers with different skill levels.

We find that the AI improves the productivity of taxi drivers by shortening the length of search time by 5%. Importantly, the productivity gain is all concentrated on the low-skilled drivers: the impact on the low-skilled drivers—where skill is defined by the past driving performance—is 10% whereas the impact on the high-skilled drivers is nearly zero. These results indicate that AI is a substitute to worker skill at least in this context, suggesting that AI may not always accompany skill-biased technological change. Rather, AI has a potential to reduce the inequality among workers within the same occupation. Nonetheless, AI did not completely eliminate the productivity gap between the high- and low-skilled drivers, implying that there is skill component of high-skilled drivers that cannot be fully replaced by AI at least at this stage of technology development. Our result also indicates that the impact of AI on employment is more nuanced and complex than the simple replacement story.

[4] For these reasons, behaviors of taxi drivers are widely studied in labor economics (e.g., Camerer et al. 1997; Farber 2005, 2008; Haggag et al. 2017)

Of course, this study is a case study and our findings only speak to the case of the taxi drivers. Nonetheless, the intersection of the skill required for the job and the capability of AI in our case is not unique to this setting. To the extent that the core skill of the jobs involves a prediction task, and the AI improves the accuracy of such prediction task, our results may be applicable to such occupations as well. For example, low-skilled paralegals might be more benefited by AI reviewing contracts for unusual clauses, and low-skilled pathologists might be more benefited by AI detecting the potential cells with malign tumors than high-skilled paralegals and pathologists (Webb 2020).

2 Background and Data

2.1 Setting

Our setting is taxi industry in Yokohama city, which is adjacent to Tokyo. Yokohama city has a population of 3.75 million, the second-largest in Japan next to Tokyo. With an area of 435 km^2 (about 7 times the area of Manhattan), Yokohama city is divided into 18 wards. There are 8,842 registered taxi drivers working for taxi operating firms in the city as of December 2019. The taxi drivers registered at Yokohama areas (Yokohama city and three other near-by cities, Kawasaki, Yokosuka, and Miura) are allowed to drop-off anywhere outside of Yokohama area but not allowed to pick-up outside of Yokohama area.[5]

The same price schedule applies to most of the taxis in the city. The fare is the sum of the fixed charge for the first 2 km (JPY740) and the variable charge after first 2 km which is determined by the distance and time, as in other usual settings of taxi.

Taxi drivers in our data works for taxi operating companies, and they are paid by the fixed percentage of the fares they collect (usually between 50% and 60%) with a guarantee of baseline salary so that they will not work below minimum wage. The drivers do not incur any cost of gas, hence they have a strong incentive to increase sales.

2.2 AI Navi

The particular AI technology we study is called "AI Navi" developed by a tech company. AI Navi is designed to help the drivers search for customers. Using machine learning technique, AI Navi's demand-forecasting skill is trained by the recent driving records in Yokohama city.

[5] Note that online ride-hailing services, such as Uber and Grab, were not permitted in Japan during our sample period.

More specifically, AI Navi, when it is turned on, suggests routes to taxi drivers when the taxi does not carry customers. The suggested routes are drawn to maximize the probability a taxi will catch customers given the location of the taxi based on the predicted demand. Thus, the AI is expected to improve the productivity of taxi drivers by reducing the time they search for customers. As mentioned earlier, more than 80% (87.3% from our data excluding breaks) of driver's working time is devoted to searching for customers, and hence improvement in search time is critical for drivers. Our interviews with some drivers reveal that there are many different strategies to search for the customers, and each driver takes his/her own strategy. Appendix Figure A1 displays the snapshot of AI Navi when it is turned on. AI Navi shows the suggested routes with a red arrow given a taxi's current location, and red dots indicate the locations with potential customers.

2.3 Data

Our data is provided by the tech company that developed "AI Navi". To gather field data, the company provided AI Navi to 5.9% (= 522/8842) of taxi drivers working for taxi operating firms in Yokohama city *for free* during December 3, 2019 to December 31, 2019. Taxi drivers who participated in this free-trial face no reward or penalty for use or non-use of the application. Thus, it is totally up to the discretion of taxi drivers whether to use it and how often to use it. In addition, we have data for the period two months before the free-trial started (i.e., October and November 2019), which we use to construct drivers' skills based on their productivity in this pre-period. Unfortunately, we do not have any information about the drivers' other characteristics such as age, gender, and tenure.

Our unit of observation is each vacant cruise, that is, a cruise during which drivers are searching for customers. Formally, we define the vacant cruise as the time between when vacant cruises starts (i.e., dropping-off the last customers) and when vacant cruise ends (i.e., finding and picking-up customers).

The original data consists of 154,444 vacant cruises in December 2019. We made the following sample restrictions. First, because AI Navi is specifically designed to help taxi drivers find the customers on the streets (not at taxi stands), we exclude the vacant cruises which end at the taxi stands (N= 58,025). Second, we exclude the vacant cruises of drivers whose pre-period data do not exist to construct our skill measures (N= 1,264). Third, following the classification by the company, vacant cruises in which a taxi did not move for more than 30 minutes is considered as being in a break, and thus are excluded (N= 5,111).[6]

[6] Similarly, Haggag et al. (2017) drop observations in which a driver has no customers for over 60 minutes

Finally, we exclude the top 1%-tile (57 min) of vacant cruise time to drop outliers (N= 900).[7] The final sample consists of 89,144 vacant cruises of 522 drivers.

Among 89,144 vacant cruises, the number of the vacant cruises when AI Navi is turned on and off are 4,772 (5.4%) and 84,372 (94.6%) respectively. In terms of the number of drivers, out of total of 522 drivers, 212 drivers (41%) used AI Navi at least once, and 310 drivers (59%) never used it during the trial period. Thus, whereas the overall utilization of AI Navi is quite low, nearly 40% of drivers at least experimented with it. We call the sample of all the drivers as "full" sample and the sample of drivers who used AI Navi at least once during the trial period as "Navi users" sample throughout the paper. Generally, our results are robust to the use of either dataset.

Appendix Figure A2 shows the distribution of vacant cruise time separately for (a) when AI is turned on, and (b) when AI is turned off. Both the mean and median of vacant cruise time are *higher* when AI is turned on than when AI is turned off; the mean(median) time when AI is on is 15.0(11.3) minutes whereas the time when AI is off is 11.3(7.8), suggesting that drivers are more likely to turn on AI when it is difficult to find customers. This selective usage of AI Navi indicates that a simple comparison of the average vacant cruise time between when AI is turned on and off is problematic because it may rather reflect the difference in the underlying demand for the taxi rather than the effect of AI Navi. We discuss how we address this selection issue in the next section.

3 Empirical strategy

3.1 Hazard model

Our empirical strategy is still comparing the vacant cruise time when AI Navi is turned on and off. As mentioned earlier, however, we cannot simply compare the average vacant cruise time between AI usage and non-usage because timing of AI Navi usage could be endogenous. Thus, we compare the vacant cruise time between when AI Navi is turned on and off *within* the same drivers by including driver FEs while controlling for rich sets of fixed effects to account for underlying demand, namely 18 ward FEs, and 696 date-hour FEs (=29 days$\times$24 hours/day). Our identifying assumption, thus, is that turning on (or off) the AI Navi is quasi-random after controlling for these sets of FEs. In the next subsection, we will show that this assumption is likely to hold.

by inferring that the driver is on a break, and is not attempting to find customers. Our results are robust to excluding 60 minutes breaks (not shown).

[7] Appendix Table A1 and A2 show that our results are barely affected by the choice of the threshold.

We estimate a hazard model to allow for a variation that AI Navi is turned on (or off) during a vacant cruise. We assume that the duration of the vacant cruise (in minutes), T, follows Weibull distribution. The survival function, $S(t) = \Pr(T > t)$, which is the probability that the drivers cannot find a customer until time t, for vacant cruise s by driver i at date-hour h in ward j is:

$$S_{ijh}(t) = \exp(-\lambda_{ijh}(t) \cdot t^p)$$

where

$$\lambda_{ijh}(t) = \exp\{-p(\alpha \cdot \text{AI Navi usage}_{ijhs,t} + \text{driver FE}_i + \text{ward FE}_j + \text{date-hour FE}_h)\}. \quad (1)$$

AI Navi usage is a dummy variable that takes the value of one when AI Navi is turned on and zero otherwise. The parameter p captures the duration dependence of the baseline hazard, where $p = 1$ implies the absence of the duration dependence, $p > 1$ ($log(p) > 0$) implies the positive duration dependence, and $p < 1$ ($log(p) < 0$) implies the negative duration dependence.

This model can be interpreted as

$$log(\text{vacant cruise time}_{ijhs}) = \alpha \cdot \text{AI Navi usage}_{ijhs,t} + \text{driver FE}_i + \text{ward FE}_j + \text{date-hour FE}_h + \epsilon_{ijhs} \quad (2)$$

where ϵ follows an extreme-value distribution. Our coefficient of interest is α which corresponds to percentage change in vacant cruise time. We test whether AI Navi usage reduces the time to find a customer ($\alpha < 0$).

To consider the effects of driver skill and demand condition, we construct the following two indices: driver skill index and vacancy index. Both indices are constructed using the vacant cruise data from October and November—a period *before* the trial period. The driver skill index is constructed in the following way. First, we estimate the hazard model of equation (1), regressing the vacant cruise time onto ward FEs, date-hour FEs, and driver FEs. Then, we flip the sign of the estimated driver FEs so that a higher skill index reflects more skilled drivers, and then standardize it to the mean of 0 with standard deviation of 1. This index essentially captures each driver's skill in searching for customers. Because our skill measure is constructed based on the worker productivity of the same drivers using past records, this index better reflects the actual skill of workers than commonly-used alternatives such as tenures and education levels. Similarly, we construct a vacancy index by estimating the same Weibull hazard model of equation (1), regressing vacant cruise time onto driver

FEs and ward-day-hour FEs. The estimated ward-day-hour FEs—which capture the average demand for a taxi at each ward at each day-hour (e.g., 10 pm on Wednesday at Ward 1)—is our vacancy index. The higher vacancy index means that it takes more time to find the customer. Thus, the higher the vacancy index, the lower the demand for a taxi at ward-day-hour level.

3.2 Credibility of underlying assumption

Recall that our identifying assumption is that turning on (or off) the AI Navi is as good as random within the same driver in similar demand conditions, that is, after controlling for ward FEs, and date-hour FEs in addition to driver FEs. To assess the plausibility of this assumption, we estimate a logistic regression, where the outcome is a dummy that takes one when AI Navi is turned on and zero otherwise, on driver skill index, vacancy index, and its interaction *with* and *without* the same sets of fixed effects as equation (1), namely, ward FEs, date-hour FEs, and driver FEs.

Table 1 shows the results. Column (1) shows that without the above-mentioned set of fixed effects, the skill index is negative ($p < 0.10$), indicating that low-skilled drivers are more likely to use AI Navi. More importantly, the vacancy index is positive and highly statistically significant ($p < 0.01$), suggesting that drivers are more likely to turn on AI Navi when the demand is low (recall that vacancy index is high when the demand is low). However, once we add ward FEs and date-hour FEs in column (2), the vacancy index is no longer statistically significant at the conventional level nor economically large. This result suggests that once we properly control the demand using ward FEs and date-hour FE, whether to turn on or off AI Navi can be arguably viewed as good as random.

Column (2) of Table 1 also shows that the interaction term of skill index and vacancy index is also not statistically significant nor economically large, which mitigates the concern that the productivity gains by the skill shown in the next section, is simply driven by differential timing of AI Navi usage (i.e., low-skilled drivers are more likely to turn on AI Navi when there is large scope to shorten the search time). Columns (3) and (4) repeat the same exercise only for the drivers who use AI Navi at least once during the trial period ("Navi users sample"). We are reassured that we find similar patterns as the full sample in columns (1) and (2). Finally, column (5) adds driver FEs to column (4), and the estimates on both vacancy index itself and interaction term of skill index and vacancy index become even smaller and reaches close to zero.

To sum, although it seems that AI Navi is more likely to be used when demand is low, such underlying demand condition can be well-controlled by including ward FEs and date-time FEs. This is plausible as none of the drivers are exposed to this application before, and thus they are likely to randomly experiment to turn it on and off especially at the beginning of the trial period. Even if our rich controls still fail to fully capture the underlying demand conditions, to the extent that drivers are likely to turn on AI Navi when the demand is low (and finding customers is difficult), the bias goes against our finding that AI Navi reduces the search time. In other words, our estimate presented in the next section may provide the lower bound in terms of the magnitude of the impact of AI.

4 Results

4.1 Overall productivity

Table 2 reports the main result of estimating equation (1). Columns (1)-(3) report the average gains in productivity measured by the reduction in search time. Column (1) shows that the AI reduces the time of finding customers by 5.1% using the full sample. We show graphically the fitness of our hazard model. Figure 1 compares the estimated survival curve from column (1) of Table 2 (solid) and the Kaplan-Meier curve (dash). The fit is reasonably high, suggesting that Weibull distribution well captures the underlying hazard.

Column (2) of Table 2 limits the sample only to AI Navi users and finds similar results.[8] Because the utilization rate of AI Navi is low, one concern is that the vacant cruises with and without turning on AI Navi within the same drivers could be unobservably different even after controlling for ward and date-hour FEs. To address this issue by ensuring sufficient overlap in characteristics between the vacant cruises with and without AI Navi usage, we trim the sample based on propensity score. Specifically, we compute the propensity score of turning on AI Navi from the logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Column (3) limits the sample to vacant cruises whose propensity score is between 0.1 and 0.9 (Imbens 2015).[9] Although the number of observations substantially decreases, it is reassuring that the estimate in column (3) is very similar to those in columns

[8] This is expected since our source of variation for identification is within drivers, and thus the drivers who never used the AI Navi (who are included in column (1) but not in column(2)) only contribute to the precision of ward FEs and date-hour FEs.

[9] Crump et al. (2009) suggest to drop the observations with the propensity score outside of the range between 0.1 and 0.9 as a close approximation of the optimal rule, and demonstrate that the rule effectively resolves the problems arising from the lack of the sufficient overlap in the observable characteristics for the wide range of distributions. This method is used for the robustness check of various empirical papers including Currie and Walker(2011) and Gibson and McKenzi (2014).

(1) and (2).[10] The estimates of $\log(p)$ are all positive, indicating that the hazard rate is increasing with respect to the vacant time.

4.2 Productivity gain by skills

Columns (4)-(6) of Table 2 report the productivity improvement by the skill level, and show that productivity gains are concentrated on the low-skilled drivers. Specifically, we divide the sample into half by median of skill index into high- and low-skilled drivers.[11] Column (4) shows that whereas the AI reduces the search time by as much as 10% for the low-skilled drivers, the corresponding gain for the high-skilled drivers is essentially zero.[12] As a result, the AI narrows the productivity gap between high- and low-skilled drivers by about 30%.[13] Nonetheless, AI did not completely eliminate the productivity gap, implying that there is unobserved skill component of high-skilled drivers that cannot be fully replaced by AI. We find similar results in columns (5) and (6), where the former limits the sample only to AI Navi users, and the latter further limits the sample only to vacant cruises with propensity scores ranging between 0.1 and 0.9. Appendix Table A3 further divides the driver skill into a quartile, and finds that the productivity gains are concentrated in the first and second quartiles, which echos the previous findings.

One remaining concern could be that even though the AI Navi would have also benefited the high-skilled drivers, they simply did not follow the navigation routes suggested by the AI. Since AI Navi assists a prediction task only but not a decision task, it is up to drivers whether to follow the AI's prediction.[14] High-skilled drivers may trust less in AI because they may have high self-confidence on their own judgements and/or they may be also more likely to spot imperfections of AI.

To test this possibility, we also control for the "Navi compliance rate", which is the fraction of navigation routes AI suggested that drivers did follow, calculated for each vacant

[10] Appendix Table A1 and A2 report the results of a) without dropping the top 1% outliers, and b) changing the cut threshold from 1% to 3%. Our estimates are not sensitive to these changes.

[11] The fraction of low-skilled drivers are similar among the Navi-users (48%= 101/211) and Non Navi-users (52%= 160/310).

[12] The differences between low- and high-skilled drivers are statistically significant at the conventional level in all columns (4)-(6).

[13] The estimate on the low-skilled drivers (0.098 from column (2) of Table 2) is divided by the sum of the average driver FEs for high- and low-skilled drivers (0.089 + 0.244). When we use median instead of mean, the gap is narrowed by 37%.

[14] Agarwal et al. (2018, 2019) consider a decision task distinct from a prediction task, where a prediction is an input to a decision task. In this framework, AI saves time and improves accuracy in generating predictions, which allows more nuanced decisions through the reduction of uncertainty in prediction.

cruise.[15] Appendix Table A4 presents the results. Odd-numbered columns replicate the results of Table 2 for ease of comparisons. Even-numbered columns add the interaction of AI Navi usage dummy and Navi compliance rate to adjacent odd-numbered columns. Whereas the interaction term is negative as expected (i.e., higher compliance reduces the search time), our coefficient of interest ("Navi usage × low-skilled") is hardly changed. Therefore, it is unlikely that the differential impact by skill level is driven by the compliance rate.

We also investigate whether the impact of AI evolves over time. Whereas our trial period is limited to one month in December, we nonetheless split the month into the first and second two weeks. Table 3 shows the results using the Navi users sample. First, column (1) shows that the AI's positive impact is immediate and observed already in the first two weeks (labeled "Navi usage"). This result is reassuring because drivers are more likely to randomly experiment to turn it on and off especially at the beginning of the trial period, mitigating the concern that the timing of switching on AI Navi could be endogenous to local demand. Second, there is some improvement in the last two weeks since the estimate on "Navi usage × 3rd/4th weeks", which captures the *additional* improvement to the first two weeks, is negative. But the magnitude is small and not statistically significant. Since this interaction term is likely to capture the combined effects of the selection (of drivers who still use the AI Navi in the last two weeks), and potential learning by doing, we limit our sample to drivers who use the AI Navi at least once in the 4th week to partially account for the selection issue. Column (2) shows that the interaction term is negative but still far from statistically significant, suggesting that learning is probably limited. An alternative interpretation is that the positive impact of AI observed in the first two weeks does not fade away in the second two weeks.

The rest of Table 3 repeats the same exercise for the low- and high-skilled drivers, separately. As for the low-skilled drivers in columns (3)-(4), we see an immediate AI's impact in the first two weeks but also do not see much learning in the last two weeks. As for the high-skilled drivers in columns (5)-(6), we do not observe any effects in the first two weeks. Whereas we see a modest effect in the last two weeks, the estimate is far from statistically significant. In sum, we do not observe much learning in this setting.[16]

Robustness. Finally, Appendix Table A5 instead reports the estimates of the Cox

[15]The average compliance rate is 53.4%. The compliance rate is in fact slightly higher for low-skilled drivers (56.2%) than high-skilled drivers (49.9%). This rate is calculated by the tech company based on their definition.

[16]Whereas we fully control the differential demand by including ward and date-hour FEs, there could be still some unobserved difference in demand between the first and second half of December, the latter of which includes the holiday season.

Proportional hazard model to allow for non-parametric baseline hazard. To make them comparable to the estimates of Weibull hazard model, Appendix Table A6 convert the estimates in Table 2 to hazard ratios. Note here that the estimates greater than one mean that the probability of picking up a customer increases as the corresponding variable increases while the estimates less than one means that the probability decreases. Comparing Appendix Tables A5 and A6, the estimates from two models are almost identical.

4.3 Another productivity measure

Thus far, we use the length of search time as a measure of productivity, and find that this AI boosts the productivity on this dimension for low-skilled drivers. However, another natural candidate of productivity measure is the sales. Whereas the reduction in search time leads to the increase in the number of rides, the fare per ride might decrease if the AI directs to the locations with customers with short rides. This might happen because AI Navi is designed to maximize the probability of catching the customers, and is *not* designed to increase the fares or sales.

We have data on fare per ride. Appendix Figure A3 shows the distribution of fare per ride separately for rides of customers who are found (a) when AI Navi is turned on, and (b) when AI Navi is turned off. Because of missing data in fares, the sample size is roughly three-fourth of the sample used in the analysis for search time in Table 2.[17] Both the mean and median of fare per ride are slightly lower for rides of customers found when AI Navi is turned on (N= 3,624) than when AI is turned off (N= 30,417); the mean(median) fare of the former is JPY1,524(1,100) whereas that of latter is JPY1,656(1,130). Note that JPY100 is roughly 1USD.

Table 4 reports the results of OLS regression on fare per ride. Specifically, we regress fare per ride on the same sets of fixed effects included in equation (1), namely, driver, ward, and date-hour FEs. Note that we include driver FEs, and thus the comparison is again within the same driver. Here, we limit the sample to AI Navi users.[18]

Column (1) of Table 4 shows that the rides of customers found when AI is turned on are slightly cheaper than the rides of customers found when AI is turned off. However, the magnitude (JPY46) is tiny compared to the average fare of JPY1500-1600 per ride. Column (2) takes the log as the outcome, and shows that the difference is around 2%. Columns (3) and (4) report the differential effect by the skill level. We find that the reduction in fare per ride is concentrated on the low-skilled drivers. Columns (3) and (4) show that for low-skilled

[17] Appendix Table A7 verifies that our main results reported in Table 2 hold in this limited sample as well.
[18] Appendix Table A8 reports the results of using the full sample, and finds very similar results.

drivers the fare per ride is reduced by JPY57 or 3.9%. Interestingly, the estimates on the logged outcome are always more precise than the outcome in levels, indicating that most of the reduction in fares comes from the increases in short-rides. We do not see any effect on the high-skilled drivers. These results suggest that AI seems to direct the low-skilled drivers to the locations with low-hanging fruits. This happens probably because this AI is designed to find customers and not to maximize sales as mentioned earlier. However, even for the low-skilled drivers, the magnitude of reduction in fare is rather small (3.9%) which does not offset the reduction in search time (10.6% from column (5) of Table 2).

5 Discussion and Conclusion

We investigate the impact of AI on worker productivity in the context of taxi drivers. The AI in our setting increases the accuracy of predicting consumer demand, and navigates the taxi drivers to the location with high demand. We find that the AI improves the worker productivity measured by the length of search time whereas all the gains are concentrated on the low-skilled drivers. This result suggests that the AI is a substitute to worker skill, and that AI may not always enhance skill-biased technological change. Rather, to the extent that productivity is reflected by wages, AI has a potential to reduce the wage inequality across workers at least within the same occupation. Our result shows that the impact of AI on employment is more nuanced and complex than the simple replacement story.

Our findings are consistent with Webb (2020) in that skilled workers are more exposed to AI than unskilled workers; the key difference is that we show that the same prediction applies to workers with differential skills *within* a single occupation where Webb (2020) shows it *across* occupations. In addition, Webb (2020) argue that AI will reduce 90:10 wage inequality under the assumption that the historical patterns of long-run substitution by robots and software will also apply to AI. Using actual field data, we show that AI indeed narrows productivity gap between high- and low-skilled workers at least in our setting.

Our result that AI adoption makes driver's demand-forecasting skill less important and obsolete have some implications for the hiring strategy of taxi operation company; the skills that cannot be automated by AI such as social skill (Deming and Kahn 2018) may become more important in hiring. Indeed, Acemoglu et al. (2020) document that establishments exposed to AI substantially changed the task contents of job openings, suggesting that the penetration of AI induced the reorganization of worker skill compositions.

This study faces several limitations. First, whereas we show that low-skilled drivers are benefited from AI, one puzzle is that the utilization rate of AI is low even among the low-skilled drivers (5.9%).[19] One possibility is that the productivity gain of 10%, which translates into the reduction of search time by 1.5 min, is not large enough to make them recognize the improvement, especially because those low-skilled drivers may be inexperienced. Relatedly, more than half of the drivers—who are given the opportunity to use the application for free—never even bother to use it. In fact, 51.6% of these never-users are low-skilled drivers who would have benefited if they had used it.[20] While this is beyond the scope of this study to identify the reasons for aversion, drivers might feel competitive pressure from AI which might replace their core skill of demand forecasting.[21] Second, we can speak little about the general equilibrium effect; what if all the drivers in the area adopt this AI technology. One concern is that taxi drivers in the area compete for the same customers, and end up engaging in business stealing if the market size stays constant. However, the consumers are benefited from the shorter time of finding a taxi. To the extent that this improved convenience stimulates the further demand for a taxi, the market could expand and the social welfare might improve.

Finally, one might wonder if our finding can be generalized beyond the case of taxi drivers. Although taxi drivers as an occupation might be completely displaced once self-driving cars with demand-forecasting AI is achieved, such drastic transformation may take time because information required for driving task such as the road environment is much less regularized than information required for demand-forecasting task such as passengers' location. As Autor (2015) points out, automating a task is much more costly under non-regularized environment than regularized environment, and the cost is likely to exceed the wage saving.[22] As a result, the partial automation of tasks within an occupation (like this taxi case) is likely to persist. Therefore, our finding might be applicable to other occupations in which the core skill involves a prediction task. Agarwal et al. (2019) classify such type of AI usage as "augmenting labor on decision tasks" where the automation of prediction through AI can improve decision-making by humans and consequently the productivity of

[19] The corresponding figure for the high-skilled drivers is 4.8%.

[20] Note that since we construct the skill index using the full sample, we can classify non-users into high- and low-skilled drivers.

[21] This may not be so-called algorithm aversion, that is, even when an algorithm consistently beats human judgment, people prefer to go with their own judgement (Dietvorst et al. 2015; 2016). Never-users have not seen that the AI does better than their own judgement. One possible reason is that the average age of drivers is high. Although we do not have information on driver characteristics at the individual level, the average age of taxi drivers at Yokohama area from the official statistics is 61.2 as of December 2019, and thus using modern application could be a challenge for a large fraction of them. In addition, information sharing among drivers seems rare according to our interviews with taxi drivers.

[22] In addition, removing completely human operations involves substantial risks, because the cost of failure (such as injuries or deaths) can be very high (Agarwal et al. 2019).

labor. For example, paralegals—affected by legal tech AI that helps them review contracts to identify unusual clause, and pathologists—affected by diagnostic imaging AI that detects malign tumors, are likely candidates. Whether our finding can be generalized to other settings is an interesting avenue for future research.

References

[1] Alekseeva, L, Azar, J., Gine, M., Samila, S., and Taska, B. (2020) "The demand for AI skills in the labor market", *CEPR Discussion Paper DP14320*.

[2] Acemoglu, D, and Restrepo, P. (2020) "Robots and Jobs: Evidence from US Labor Markets", *Journal of Political Economy*, 128(6): 2188-2244.

[3] Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2020) "AI and Jobs: Evidence from Online Vacancies", *NBER Working Paper No. 28257*.

[4] Agrawal, A., Gans, J. S. , and Goldfarb, A. (2018) *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Review.

[5] Agrawal, A., Gans, J. S. , and Goldfarb, A. (2019) "Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction'", *Journal of Economic Perspectives*, 33(2): 31–50.

[6] Autor, D. H., Levy, F., and Murnane, R. J. (2003) "The skill content of recent technological change: An empirical exploration", *The Quarterly Journal of Economics*, 118(4): 1279-1333.

[7] Autor, D. H., Katz, L. F., and Kearney, M. S. (2008) "Trends in U.S. Wage Inequality: Revising the Revisionists", *The Review of Economics and Statistics*, 90(2): 300–323.

[8] Autor, D H. (2015), "Why Are There Still So Many Jobs? The History and Future of Workplace Automation", *Journal of Economic Perspectives*, 29(3): 3-30.

[9] Bartel, A., Ichniowski, C., and Shaw, K. (2007) "How does information technology affect productivity? Plant-level comparisons of product innovation, process improvement, and worker skills", *The Quarterly Journal of Economic*, 122(4): 1721-1758.

[10] Brynjolfsson, E., Rock, D., and Syverson, C. (2018) "Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics", in *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press.

[11] Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997) "Labor Supply of New York City Cabdrivers: One Day at a Time", *Quarterly Journal of Economics*, 112(2): 407–441.

[12] Card, D., and Lemieux, T. (2001), "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis", *The Quarterly Journal of Economics*, 116(2): 705–746,

[13] Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009) "Dealing with limited overlap in estimation of average treatment effects", *Biometrika*, 96(1): 187-199.

[14] Currie, J., and Walker, R. (2011) "Traffic Congestion and Infant Health: Evidence from E-Z Pass", *American Economic Journal: Applied Economics*, 3(1): 65–90.

[15] Deming, D., and Kahn, L. B. (2018) "Skill Requirements across Firms and Labor Markets: Evidence from Job Postings for Professionals", *Journal of Labor Economics*, 36(S1): S337-S369.

[16] 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.

[17] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2016) "Overcoming Algorithm Aversion: People will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them", *Management Science*, 64(3): 1155-1170.

[18] Farber, H.S. (2005) "Is Tomorrow Another Day? The Labor Supply of New York City Cabdrivers.", *Journal of Political Economy*, 113(1): 46–82.

[19] Farber, H.S. (2008) "Reference-Dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers", *American Economic Review*, 98(3): 1069-1082.

[20] Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., Wang, D., Youn, H., and Rahwan, I. (2019), "Toward understanding the impact of artificial intelligence on labor", *Proceedings of the National Academy of Sciences*, 116: 6531-6539.

[21] Felten, E. W., Raj, M., and Seamans, R. (2018) "A Method to Link Advances in Artificial Intelligence to Occupational Abilities", *AEA Papers and Proceedings*, 108: 54-57.

[22] Felten, E. W., Raj, M., and Seamans, R. (2019) "The occupational impact of artificial intelligence: Labor, skills, and polarization", *SSRN working paper No. 3368605.*

[23] Grenna, J., and Michaely, R. (2020) "Artificial Intelligence and High-Skilled Work: Evidence from Analysts", *Swiss Finance Institute Research Paper No. 20-84.*

[24] Gibson, J., and McKenzie, D. (2014) "The Development Impact of a Best Practice Seasonal Worker Policy", *The Review of Economics and Statistics*, 96(2): 229–243.

[25] Haggag, K., McManus, B., and Paci, G. (2017) "Learning by Driving: Productivity Improvements by New York City Taxi Drivers", *American Economic Journal: Applied Economics*, 9(1): 70-95.

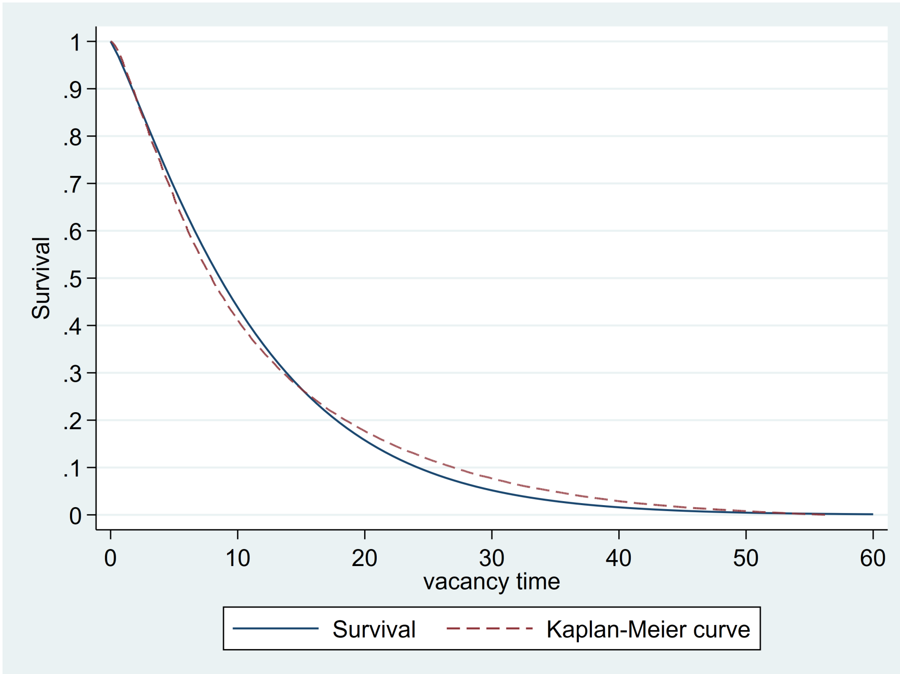
[26] Imbens G. W. (2015) "Matching methods in practice: Three examples", *Journal of Human Resources*, 50(2): 373-419.

[27] Katz, L. F., and Murphy, K. M. (1992) "Changes in Relative Wages, 1963-1987: Supply and Demand Factors", *The Quarterly Journal of Economics*, 107(1): 35-78.

[28] Webb, M. (2020) "The impact of artificial intelligence on the labor market", *SSRN working paper No. 3482150.*

Figures and tables

Figure 1: Model Prediction vs. Kaplan-Meier Curve



Notes: The figures compares the estimated survival curve from column (1) of Table 2 (solid) and the Kaplan-Meier curve (dash).

Table 1: Logistic Regression for AI Navi Usage

	(1) Full	(2) Full	(3) Navi users	(4) Navi users	(5) Navi users
skill index	-0.209*	-0.248*	-0.281**	-0.281***	
	(0.125)	(0.127)	(0.110)	(0.109)	
vacancy index	0.198***	-0.046	0.185***	-0.055	0.018
	(0.031)	(0.039)	(0.032)	(0.037)	(0.040)
skill index × vacancy index	0.031	0.024	0.021	-0.000	0.010
	(0.029)	(0.040)	(0.033)	(0.038)	(0.037)
driver FE					✓
ward FE		✓		✓	✓
date-hour FE		✓		✓	✓
N	88,864	83,355	46,046	43,476	43,476
N of drivers	522	520	212	212	212
Log-likelihood	-18,442	-16,547	-15,111	-13,383	-8,247

Notes: The outcome is an AI Navi usage, which is a dummy that takes one when AI Navi is turned on. Standard errors clustered on drivers are reported in parentheses. The higher skill index indicates more skilled drivers. The higher vacancy index indicates less demand for taxi at ward-day-hour level. “Full” in columns (1)-(2) is the sample of all drivers, and “Navi users” in columns (3)-(5) is the sample of drivers who used AI Navi at least once during the trial period. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table 2: Weibull Hazard Regression

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	-0.051***	-0.056***	-0.062***			
	(0.018)	(0.018)	(0.021)			
Navi usage × low-skilled				-0.098***	-0.107***	-0.096***
				(0.024)	(0.023)	(0.029)
Navi usage × high-skilled				-0.008	-0.000	-0.022
				(0.025)	(0.026)	(0.032)
$\log(p)$	0.151***	0.141***	0.197***	0.151***	0.141***	0.198***
	(0.004)	(0.005)	(0.008)	(0.004)	(0.005)	(0.008)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	89,144	46,208	10,076	89,144	46,208	10,076
N of drivers	522	212	161	522	212	161
Log-likelihood	-125,536	-65,533	-13,811	-125,532	-65,527	-13,809

Notes: High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in columns (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table 3: Weibull Hazard Regression: Evolution of AI's Impact

	(1)	(2)	(3)	(4)	(5)	(6)
	Navi users	4th week Navi users	Navi users	4th week Navi users	Navi users	4th week Navi users
	-	-	low-skilled	low-skilled	high-skilled	high-skilled
Navi usage	-0.046**	-0.066**	-0.095***	-0.107***	0.003	-0.024
	(0.022)	(0.029)	(0.028)	(0.040)	(0.031)	(0.041)
Navi usage $\times$ 3rd/4th weeks	-0.028	-0.020	-0.011	-0.022	-0.045	-0.045
	(0.027)	(0.035)	(0.034)	(0.045)	(0.043)	(0.059)
$\log(p)$	0.141***	0.145***	0.163***	0.178***	0.141***	0.170***
	(0.005)	(0.008)	(0.008)	(0.010)	(0.007)	(0.012)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	46,208	17,990	22,549	9,021	23,659	8,969
N of drivers	212	82	106	42	106	40
Log-likelihood	-65,532	-25,397	-31,587	-12,492	-33,512	-12,464

Notes: "Navi users" sample is used. High- and low-skilled are drivers whose skill index is above median and below median, respectively. "Navi usage" captures the impact of AI in the first two weeks, and "Navi usage $\times$ 3rd/4th weeks" captures the *additional* impact of AI in the last two weeks. Columns (2), (4) and (6) limit the sample to drivers who use the AI Navi at least once in the 4th week. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table 4: OLS Regression on Fare per Ride

	(1)	(2)	(3)	(4)	(5)	(6)
	Navi users	Navi users	Navi users	Navi users	Navi users	Navi users
					$0.1 \leq PS \leq 0.9$	$0.1 \leq PS \leq 0.9$
	fare	log(fare)	fare	log(fare)	fare	log(fare)
Navi usage	-46.019	-0.022*				
	(32.150)	(0.011)				
Navi usage × low-skilled			-57.424	-0.039***	-47.225	-0.036**
			(41.307)	(0.013)	(44.002)	(0.014)
Navi usage × high-skilled			-33.117	-0.001	-78.381	-0.016
			(48.120)	(0.018)	(56.107)	(0.019)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	34,043	34,043	34,043	34,043	8,155	8,155
N of drivers	157	157	157	157	125	125
adj. R^2	0.083	0.104	0.083	0.104	0.076	0.094

Notes: "Navi users" sample is used. The outcome is level or log of fare per ride in JPY. JPY100 is roughly 1USD. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. Columns (5) and (6) further limits the sample to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. The higher skill index indicates more skilled drivers. ***, ** and * denote 10%, 5% and 1% significance level respectively.

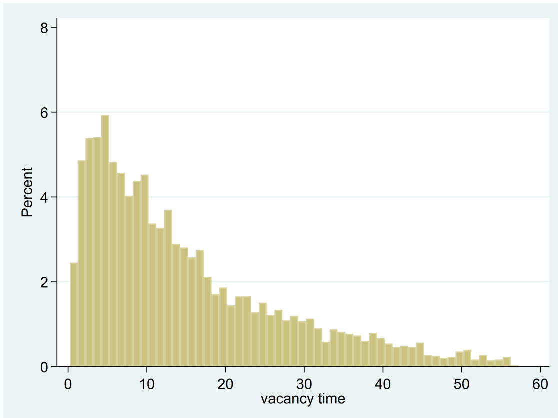
Additional figures and tables

(Not for Publication)

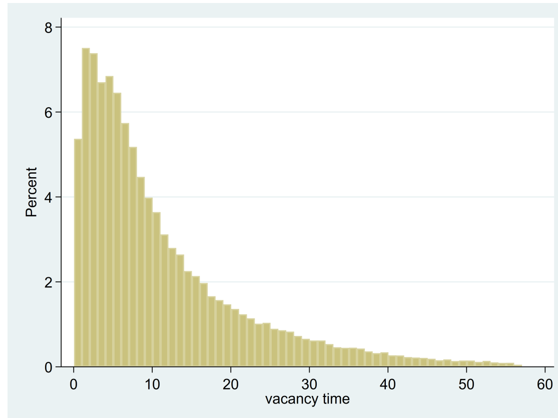
Figure A1: Snapshot of AI Navi

Notes: The figure displays the snapshot of AI Navi when it is turned on. AI Navi shows the suggested routes in green with a red arrow given a taxi's current location, and red dots indicate the locations with potential customers. © Zenrin © Mapbox

Figure A2: Histogram of Vacant Cruise Time: AI is Turned On/Off



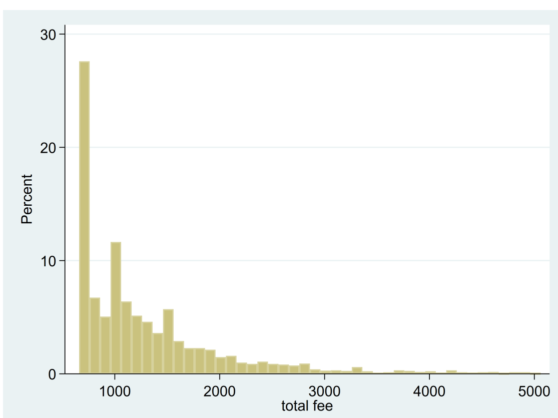
(a) AI is turned on



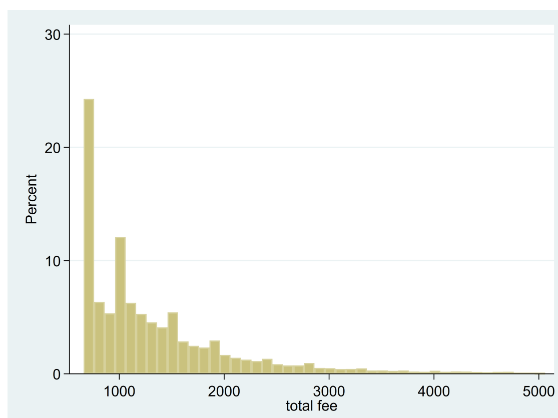
(b) AI is turned off

Note: These histograms show the distribution of vacant cruise time in the sample period (a) when AI is turned on and (b) when AI is turned off separately. The mean(median) time (a) when AI is turned on is 15.0(11.3) minutes whereas (b) when AI is turned off is 11.3(7.8). Number of observations for (a) and (b) are 4,772 and 84,372.

Figure A3: Histogram of Fare per Ride: AI is Turned On/Off



(a) AI is turned on



(b) AI is turned off

Note: These histograms show the distribution of fare per ride (in JPY) in the sample period for customers who are find (a) when AI is turned on and (b) when AI is turned off separately. JPY100 is roughly 1USD. The mean(median) fare per ride (a) when AI is turned on is JPY1,524(1,100) whereas (b) when AI is turned off is JPY1,656(1,130). Number of observations for (a) and (b) are 3,624 and 30,417.

Table A1: Weibull Hazard Regression: Without Dropping the Top 1% Outliers

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	-0.058***	-0.062***	-0.068***			
	(0.019)	(0.020)	(0.024)			
Navi usage × low-skilled				-0.085***	-0.092***	-0.087***
				(0.027)	(0.027)	(0.033)
Navi usage × high-skilled				-0.030	-0.030	-0.047
				(0.027)	(0.028)	(0.034)
$\log(p)$	0.118***	0.110***	0.165***	0.118***	0.110***	0.165***
	(0.004)	(0.005)	(0.008)	(0.004)	(0.005)	(0.008)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	90,044	47,027	10,305	90,044	47,027	10,305
N of drivers	522	214	164	522	214	164
Log-likelihood	-129,051	-67,801	-14,391	-129,050	-67,800	-14,390

Notes: We did not drop the top 1% outliers. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in columns (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A2: Weibull Hazard Regression: Dropping the Top 3% Outliers

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	-0.036**	-0.040**	-0.048**			
	(0.017)	(0.017)	(0.020)			
Navi usage × low-skilled				-0.093***	-0.089***	-0.081***
				(0.022)	(0.023)	(0.030)
Navi usage × high-skilled				0.020	0.013	-0.010
				(0.023)	(0.023)	(0.026)
$\log(p)$	0.190***	0.178***	0.238***	0.190***	0.178***	0.238***
	(0.005)	(0.005)	(0.009)	(0.005)	(0.005)	(0.009)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	87,343	44,601	9,745	87,343	44,601	9,745
N of drivers	522	209	159	522	209	159
Log-likelihood	-120,167	-61,910	-13,043	-120,160	-61,905	-13,041

Notes: We dropped the top 3% outliers. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. “Full” in columns (1) and (4) is the sample of all drivers, and “Navi users” in columns (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the “Navi users” to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A3: Weibull Hazard Regression: Skill Index Quartiles

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$
Navi usage × skill index 1st quartile	-0.090***	-0.105***	-0.091***
	(0.027)	(0.026)	(0.035)
Navi usage × skill index 2nd quartile	-0.105***	-0.109***	-0.100**
	(0.036)	(0.034)	(0.046)
Navi usage × skill index 3rd quartile	-0.028	-0.012	-0.040
	(0.030)	(0.034)	(0.047)
Navi usage × skill index 4th quartile	0.024	0.015	-0.002
	(0.040)	(0.038)	(0.038)
$\log(p)$	0.151***	0.141***	0.197***
	(0.004)	(0.005)	(0.008)
driver FE	✓	✓	✓
ward FE	✓	✓	✓
date-hour FE	✓	✓	✓
N	89,144	46,208	10,076
N of drivers	522	212	161
Log-likelihood	-125,531	-65,527	-13,809

Notes: We divide drivers into quartile by the distribution of skill index. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in column (1) is the sample of all drivers, and "Navi users" in column (2) is the sample of drivers who used AI Navi at least once during the trial period. Column (3) further limits the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A4: Weibull Hazard Regression: Controlling for Navi Compliance Rate

	(1) Full	(2) Full	(3) Navi users	(4) Navi users	(5) Navi users $0.1 \leq PS \leq 0.9$	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage × low-skilled	-0.098***	-0.100***	-0.107***	-0.108***	-0.096***	-0.094***
	(0.024)	(0.025)	(0.023)	(0.023)	(0.029)	(0.029)
Navi usage × high-skilled	-0.008	-0.013	0.000	-0.004	-0.022	-0.023
	(0.025)	(0.025)	(0.026)	(0.026)	(0.032)	(0.032)
Navi usage × Navi compliance rate		-0.066		-0.055		-0.081
		(0.046)		(0.046)		(0.049)
$\log(p)$	0.151***	0.151***	0.141***	0.141***	0.198***	0.199***
	(0.004)	(0.004)	(0.005)	(0.005)	(0.008)	(0.008)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	89,144	89,144	46,208	46,208	10,076	10,076
N of drivers	522	522	212	212	161	161
Log-likelihood	-125,532	-125,531	-65,527	-65,526	-13,809	-13,808

Notes: "Navi compliance rate" is the fraction of routes AI suggested that drivers did follow, calculated for each vacant cruise. Odd-numbered columns replicate the results of Table 2 for the ease of comparison. Even-numbered columns add the interaction of AI Navi usage dummy and Navi compliance rate to odd-numbered columns. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (2) is the sample of all drivers, and "Navi users" in columns (3) and (4) is the sample of drivers who used AI Navi at least once during the trial period. Columns (5) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A5: Cox Proportional Hazard Regression

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	1.065***	1.070***	1.076***			
	(0.022)	(0.022)	(0.027)			
Navi usage × low-skilled				1.121***	1.130***	1.116***
				(0.031)	(0.029)	(0.040)
Navi usage × high-skilled				1.016	1.007	1.031
				(0.029)	(0.030)	(0.039)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	89,144	46,208	10,076	89,144	46,208	10,076
N of drivers	522	212	161	522	212	161
Log-likelihood	-923,129	-448,085	-81,983	-923,126	-448,080	-81,981

Notes: Cox Proportional hazard regression is estimated. The estimates here indicate hazard ratios, and the estimates greater than one mean that the probability of picking up a customer increases as the corresponding variable increases while the estimates less than one mean that the probability decreases. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A6: Weibull Hazard Regression: Hazard Ratio

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	1.061***	1.066***	1.078***			
	(0.022)	(0.022)	(0.028)			
Navi usage × low-skilled				1.121***	1.131***	1.124***
				(0.031)	(0.029)	(0.039)
Navi usage × high-skilled				1.009	1.000	1.028
				(0.029)	(0.030)	(0.040)
$\log(p)$	0.151***	0.141***	0.197***	0.151***	0.141***	0.198***
	(0.004)	(0.005)	(0.008)	(0.004)	(0.005)	(0.008)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	89,144	46,208	10,076	89,144	46,208	10,076
N of drivers	522	212	161	522	212	161
Log-likelihood	-125,536	-65,533	-13,811	-125,532	-65,527	-13,809

Notes: This table converts the results from Table 2 into hazard ratio so that they are comparable to the estimates from Cox Proportional hazard regression reported in A5. Since the estimates here indicate hazard ratios, the estimates greater than one mean that the probability of picking up a customer increases as the corresponding variable increases while the estimates less than one mean that the probability decreases. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A7: Weibull Hazard Regression: Fare Data Existing Sample

	(1) Full	(2) Navi users	(3) Navi users $0.1 \leq PS \leq 0.9$	(4) Full	(5) Navi users	(6) Navi users $0.1 \leq PS \leq 0.9$
Navi usage	-0.065***	-0.062***	-0.068***			
	(0.021)	(0.021)	(0.024)			
Navi usage × low-skilled				-0.106***	-0.111***	-0.097***
				(0.028)	(0.027)	(0.034)
Navi usage × high-skilled				-0.028	-0.006	-0.032
				(0.029)	(0.030)	(0.035)
log(p)	0.154***	0.146***	0.212***	0.154***	0.146***	0.213***
	(0.005)	(0.006)	(0.009)	(0.005)	(0.006)	(0.010)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	70,171	34,043	8,155	70,171	34,043	8,155
N of drivers	420	157	125	420	157	125
Log-likelihood	-98,439	-48,060	-11,058	-98,436	-48,055	-11,057

Notes: The sample is limited to vacant cruises with fare data available. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in columns (2) and (5) is the sample of drivers who used AI Navi at least once during the trial period. Columns (3) and (6) further limit the "Navi users" to vacant cruises whose propensity score (PS) is between 0.1 and 0.9. PS is computed by predicting the probability after logistic regression of AI Navi usage dummy on driver, ward, and date-hour FEs. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5% and 1% significance level respectively.

Table A8: OLS Regression on Fare per Ride: Full Sample

	(1) Full fare	(2) Full log(fare)	(3) Full fare	(4) Full log(fare)
Navi usage	-44.841	-0.021*		
	(29.953)	(0.011)		
Navi usage × low-skilled			-57.212	-0.040***
			(38.023)	(0.013)
Navi usage × high-skilled			-30.802	0.000
			(46.702)	(0.018)
driver FE	✓	✓	✓	✓
ward FE	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓
N	70,171	70,171	70,171	70,171
N of drivers	420	420	420	420
adj. R^2	0.091	0.118	0.091	0.118

Notes: "Full sample" is used. The outcome is level or log of fare per ride in JPY. JPY100 is roughly 1USD. High- and low-skilled are a dummy for drivers whose skill index is above median and below median, respectively. Standard errors clustered on drivers are reported in parentheses. The higher skill index indicates more skilled drivers. ***, ** and * denote 10%, 5% and 1% significance level respectively.