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

July 5, 2021

Abstract

Past studies of Artificial Intelligence (AI) on employment examine its impacts *across* occupations and implicitly assume that all workers in the occupation are *uniformly* affected by AI, ignoring the substantial *within*-occupation skill heterogeneity. We study the impact of a demand-forecasting AI on productivity in the context of taxi drivers. We find that AI improves drivers' productivity by shortening the search time for customers by 5%, on average. Importantly, the productivity gain is entirely accrued by low-skilled drivers, whereas the corresponding gain for high-skilled drivers is nearly zero, suggesting that AI is a substitute for skill, at least in this context.

Keywords: Artificial Intelligence, Skill, Productivity, Taxi-drivers, Prediction, Demand forecasting, Machine learning
JEL codes: J22, J24, L92, R41

*We thank Kensuke Teshima, Hongliang Zhang, and seminar participants at Hitotsubashi University, Monash University, Nihon University, and Osaka University for their excellent comments and suggestions. The data used in this paper are provided by an anonymous tech company that developed the AI application. We have no financial support 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 the potential to drastically reshape employment (Brynjolfsson et al. 2018).[1] The impact of AI on employment, however, could be fundamentally different from those of the past technologies, including information technology (IT) and robotics, which are considered to be skill augmenting and inequality enhancing.[2] Whereas past technologies have replaced routine and manual tasks, AI replaces 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 studies examining the impact of AI consider the differential impact *across* occupations based on the degree to which each occupation is exposed to AI.[3] As a result, these studies implicitly assume that all workers *within* the occupation are *uniformly* affected by AI.[4] Even within an occupation, however, 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 levels of skill within an occupation provides a deeper understanding of AI's complex impact on human labor at a granular level. Such evidence from micro-level data is scant.[5]

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 two reasons. First, a worker's individual productivity is easily measured by the length of time it takes to search for customers as well as sales; each driver works independently and has considerable discretion as to how they search for customers. Our data show that over 80% of drivers' working time is devoted to searching for (and the remaining to carrying) customers, excluding breaks. Thus, searching for customers is among the most important tasks for taxi drivers, and the length of search time is an important productivity measure. Second, the work environment of taxi drivers offers a clean setting to study productivity, because all drivers work in a very similar setting; taxi drivers possess the same capital and charge the same prices, in the absence of the other confounding factors of productivity gain, such as

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

[2] E.g., Autor et al. (2003); Bartel et al. (2007); Acemoglu and Restrepo (2020).

[3] E.g., Felten et al. (2018, 2019); Frank et al. (2019); Webb (2020); Alekseeva et al. (2020).

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

[5] The only exception, to our knowledge, is Grennan and Michaely's (2020) study, which documents that security analysts who cover stocks that are more exposed to AI are more likely to leave the profession, and accurate analysts are more affected. The primary difference is that we study the impact on productivity.

differential input prices.[6]

The particular AI we study is called "AI Navi," which helps drivers search for customers when a taxi is vacant. The AI suggests routes based on the predicted demand to maximize the probability of catching customers, given the current location. Thus, this AI is expected to improve the productivity of drivers by reducing search time. This type of AI, which increases the accuracy of *prediction* tasks using machine learning techniques, is widely used in real business settings (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 this demand-forecasting skill is an important component of taxi drivers' skill set, the impact of AI may differ by drivers' skill.

Our empirical strategy compares the hazard rate of finding customers when AI is turned on and when it is turned off *within* the same drivers. To address the potential endogeneity of the timing of AI usage, we control for rich sets of fixed effects to account for underlying demand, namely ward, and date-hour FEs. Our identifying assumption, thus, is that turning on (or off) the AI is quasi-random after controlling for these sets of FEs. We examine the validity of this assumption empirically and find that although it seems that AI is more likely to be used when demand is low, this underlying demand condition can be well-controlled by including ward and date-hour FEs.

We find that the AI improves the productivity of taxi drivers by shortening the length of search time by 5%, on average. Importantly, the productivity gain is concentrated on low-skilled drivers: The impact on low-skilled drivers—where skill is defined by past driving performance—is 10%, whereas the impact on high-skilled drivers is nearly zero. As a result, the AI narrows the productivity gap between high- and low-skilled drivers by about 30%. These results indicate that AI is a substitute for worker skill, at least in this context. Our result indicates that the impact of AI on employment is more nuanced and complex than a simple replacement story.

Of course, this study is a case study, and our findings speak only to the case of taxi drivers. Nonetheless, the intersection of the skill required for the job and the capability of AI is not unique to our setting. To the extent that the core skill of other jobs involves a prediction task, and the AI improves the accuracy of this prediction task, our results may be applicable to other 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 within malignant tumors than

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

high-skilled paralegals and pathologists.

2 Background and Data

2.1 Setting

We study taxi drivers 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 in Yokohama areas (Yokohama city and three other nearby cities: Kawasaki, Yokosuka, and Miura) are allowed to drop-off but not allowed to pick-up anywhere outside of the Yokohama area. Note that online ride-hailing services (e.g., Uber and Grab) were not permitted in Japan during our sample period.

The same price schedule applies to most taxis in the city. The fare is the sum of the fixed charge for the first 2 km (JPY740; JPY100 is roughly 1USD) and the variable charge after first 2 km, which is determined by the distance and time, as in other usual settings of taxi transportation.

Taxi drivers in our data work for taxi-operating firms, and they are paid a fixed percentage of the fares they collect (usually between 45-60%), with a guaranteed baseline salary so that they will not work below minimum wage. Drivers do not incur any variable costs (including gas), and hence they have a strong incentive to increase sales.

2.2 AI Navi

AI Navi, developed by a tech company, is designed to help the drivers search for customers when the taxi is vacant. Using a machine-learning technique, AI Navi's demand-forecasting skill is trained by recent driving records in Yokohama city.

More specifically, AI Navi, when it is turned on, suggests routes to taxi drivers to maximize the probability a taxi will catch customers, given the location of the taxi. Thus, the AI is expected to improve the productivity of taxi drivers by reducing the search time. As mentioned earlier, over 80% (87.3% from our data excluding breaks) of a driver's working time is devoted to searching for customers, and hence improvement in search time is critical for drivers. Our interviews with drivers reveal that there are many different strategies to search for the customers, and each driver adopts his/her own strategy. Figure A1 displays a 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.

2.3 Data

Our data are 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 the period of December 3 to 31, 2019. Taxi drivers who participated in this free trial received no reward or penalty for use or non-use of the application. Thus, it was totally up to the discretion of the 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 (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 a vacant cruise as the time between when a vacant cruise starts (i.e., dropping off the last customer) and when it ends (i.e., picking-up the next customer) on the street.

The original data consist of 96,419 vacant cruises in December 2019. We make the following sample restrictions. First, we exclude the vacant cruises of drivers whose pre-period data do not exist to construct our skill measures (N = 1,264). Second, following the company's classification, vacant cruises in which a taxi did not move for over 30 minutes are considered as being on a break, and thus are excluded (N = 5,111).[7] Finally, we exclude the top 1%-tile (57 minutes) of vacant cruise time to drop outliers (N = 900).[8] The final sample consists of 89,144 vacant cruises of 522 drivers.

Among 89,144 vacant cruises, the numbers of the vacant cruises when AI 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 a total of 522 drivers, 212 (41%) used AI at least once and 310 (59%) never used it during the trial period. Thus, whereas the overall utilization of AI is quite low, nearly 40% of drivers at least experimented with it. We call the sample of all drivers the "full sample" and the sample of drivers who used AI at least once during the trial period the "Navi users sample" throughout the paper. Generally, our results are robust to the use of either dataset.

[7] Similarly, Haggag et al. (2017) drop observations in which a driver has no customers for over 60 minutes by inferring that the driver is on a break. Our results are robust to excluding 60-minute breaks (not shown).

[8] Tables A1 and A2 report the results 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.

Figure A2 shows the distributions 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 it is turned off; the mean(median) time when AI is on is 15.0(11.3) minutes, whereas the time when it 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 indicates that a simple comparison of the average vacant cruise time between when AI is turned on and when it is turned off is problematic, because it instead reflects the difference in the underlying demand for the taxi rather than the effect of AI. 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 is turned on and when it is turned off. As mentioned earlier, however, we cannot simply compare the average vacant cruise time between AI usage and non-usage, because the timing of AI usage could be endogenous. Thus, we compare the vacant cruise time between when AI is turned on and when it is turned off *within* the same drivers by including driver FE while controlling for rich sets of fixed effects to account for underlying demand, namely 18 ward FE, and 696 date-hour FE (=29 days×24 hours/day). Our identifying assumption, thus, is that turning on (or off) the AI is quasi-random after controlling for these sets of FEs. In the next subsection, we will examine the validity of this assumption empirically.

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

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

where

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

AI usage is a dummy variable that takes the value of one when AI is turned on, and zero

otherwise. The incidental parameter problems due to estimating numerous fixed effects in a non-linear model is addressed in Appendix Section C, and finds that the bias is small. 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

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

where ϵ follows an extreme-value distribution (Wooldridge 2010, p. 998). Our coefficient of interest is α, which corresponds to percentage change in vacant cruise time. We test whether AI usage reduces the time to find a customer ($\alpha < 0$).

To consider the effects of driver skill and the demand condition, we construct two indices: a driver skill index and a 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 driver, ward, and date-hour FEs. Then, we flip the sign of the estimated driver FE, 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 tenure and education level.[9] Importantly, since this skill measure captures not only the demand-forecasting skill but also other skills necessary to shorten the search time, such as driving skills[10], it is not obvious *a priori* whether AI is more beneficial for more or less skilled drivers, respectively. Similarly, we construct a vacancy index by estimating the same hazard model of equation (1), regressing vacant cruise time onto driver and ward-day-hour FEs. The estimated ward-day-hour FE—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. A higher vacancy index indicates more time to find the customer. Thus, the higher the vacancy index, the lower is the demand for a taxi at the ward-day-hour level.

[9] E.g., Katz and Murphy (1992); Card and Lemieux (2001); Autor et al. (2008).

[10] Driving skills include skills to stop in front of other cars at red traffic lights to be more visible to customers and avoid driving immediately after another vacant taxi, as doing so lowers probability of catching customers.

3.2 Credibility of the underlying assumption

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

Table 1 shows the results. Column (1) shows that without the above-mentioned set of FEs, 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 when the demand is low (recall that vacancy index is high when the demand is low). Once we add ward and date-hour FEs in column (2), however, the vacancy index is no longer statistically significant nor is it economically large. This result suggests that once we properly control the demand using ward FEs and date-hour FEs, whether to turn on or off AI is arguably considered as good as random.

Column (2) of Table 1 shows that the interaction term of the skill index and the vacancy index is also neither 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 usage (i.e., low-skilled drivers are more likely to turn on AI when the demand is low). Columns (3) and (4) repeat the same exercise only for the drivers who use AI at least once during the trial period ("Navi users sample"). We are reassured that we find similar patterns as those in the full sample in columns (1) and (2). Finally, column (5) adds driver FE to column (4), and the estimates on both the vacancy index itself and the interaction term of the skill index and the vacancy index become even smaller and reach close to zero.

In sum, although it seems that AI is more likely to be used when demand is low, this underlying demand condition can be well-controlled by including ward and date-hour FEs. This is plausible, because none of the drivers had been exposed to this application before, and thus they were likely to randomly experiment with turning 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 when the demand is low (and finding customers is difficult), the bias goes against our finding that AI reduces search time. In other words, our estimate presented in the next section may provide

the lower bound in terms of the magnitude of the AI's impact.

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 the Weibull distribution well captures the underlying hazard.

Column (2) of Table 2 limits the sample to AI Navi users and finds similar results.[11] Because the utilization rate of AI is low, one concern is that the vacant cruises with and without turning on AI within the same drivers could be different yet unobservablely, even after controlling for ward and date-hour FEs. To address this issue by ensuring a sufficient overlap in characteristics between the vacant cruises with and without AI usage, we trim the sample based on propensity score. Specifically, we compute the propensity score of turning on AI from the logistic regression of the AI usage dummy on driver, ward, and date-hour FEs. Column (3) limits the sample to vacant cruises whose propensity score lies between 0.1 and 0.9 (Imbens 2015).[12] Although the number of observations substantially decreases, it is reassuring that the estimate in column (3) is very similar to those in columns (1) and (2). 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 skill level and show that productivity gains are concentrated on low-skilled drivers. Specifically, we divide the

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

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

sample into half by the median of the skill index into high- and low-skilled drivers.[13] Column (4) shows that whereas the AI reduces the search time by as much as 10% for low-skilled drivers, the corresponding gain for high-skilled drivers is essentially zero.[14] As a result, the AI narrows the productivity gap between high- and low-skilled drivers by 27%.[15] Nonetheless, AI did not completely eliminate the productivity gap, implying that there is an unobserved skill component of high-skilled drivers that cannot be fully replaced by AI, at least at this stage of technological development. We find similar results in columns (5) and (6), where the former limits the sample to only AI users, and the latter further limits the sample to only vacant cruises with propensity scores ranging between 0.1 and 0.9. Table A3 further divides driver skill into quartiles and finds that productivity gains are concentrated in the first and second quartiles, which echoes the previous findings. Finally, Appendix Section B reports the estimates of the Cox Proportional hazard model and finds that the results are similar to those of the Weibull hazard model.

Compliance.— One remaining concern could be that even though the AI also would have benefited high-skilled drivers, these drivers simply did not follow the navigation routes suggested by the AI. Since AI Navi assists only a prediction task but not a decision task, it is up to drivers to decide whether to follow the AI's prediction.[16] High-skilled drivers may trust AI less because they may have high self-confidence in their own judgments and/or they may be more likely to spot the imperfections of AI.

To test this possibility, we also control for the "Navi compliance rate," which is the fraction of the AI's suggested routes that drivers did follow, calculated for each vacant cruise.[17] Table A4 presents the results. Odd-numbered columns replicate Table 2 for ease of comparison. Even-numbered columns add the interaction of the AI usage dummy and the 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 × high-skilled" and "Navi usage × low-skilled") are hardly changed.

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

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

[15]The difference of estimates between low- and high-skilled drivers ((-0.098-(-0.008)) from column (4) of Table 2) is divided by the difference of the average driver FE for low- and high-skilled drivers (0.244-(-0.089)). When we use the median instead of the mean, the gap is narrowed by 34%.

[16]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 by reducing uncertainty in predictions.

[17]The average compliance rate is 53.4%. The compliance rate is 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.

Therefore, it is unlikely that our results are driven by the compliance rate.

The AI's impact over time.— We also investigate whether the impact of AI evolves over time. We split the month in December 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 with turning it on and off at the beginning of the trial period, mitigating the concern that the timing of switching on AI could be endogenous to local demand. Second, there is some improvement in the last two weeks (labeled "Navi usage × 3rd/4th weeks,") which captures the *additional* improvement to the first two weeks. But the magnitude is small and not statistically significant. Since this interaction term captures the combined effects of the selection (of drivers who still use the AI in the last two weeks), and potential learning by doing, we limit our sample to drivers who use the AI at least once in the 4th week to partially account for the selection issue. Column (2) shows that the interaction term is 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 the AI's immediate impact in the first two weeks but also do not see much learning in the last two weeks. We do not observe any significant effects on high-skilled drivers in columns (5)-(6). In sum, we do not observe much learning in this setting.

4.3 Another productivity measure

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

We have data on fare per ride. Figure A3 shows the distribution of fare per ride separately for rides of customers who are found (a) when AI is turned on, and (b) when AI is turned off. Because of missing data in fares, the sample size is roughly three-fourths of the sample

used in Table 2.[18] Both the mean and the median of fare per ride are slightly lower for rides of customers found (a) when AI is turned on (N= 3,624) than (b) 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).

Table 4 reports the results of the OLS regression on fare per ride. Specifically, we regress fare per ride on the same sets of fixed effects included in equation (1), namely, ward, date-hour, and driver FEs. Note that we include driver FE, and thus the comparison is again within the same driver. Here, we use the Navi users sample.[19]

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. The magnitude (JPY46) is tiny, however, when 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 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 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 increases in short-rides. We do not see any effect on high-skilled drivers. These results suggest that AI seems to direct the low-skilled drivers to the locations with low-hanging fruit. This probably happens because this AI is not designed to maximize fares, as mentioned earlier. Even for the low-skilled drivers, however, 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 taxi drivers to locations with high demand. We find that the AI improves worker productivity measured by the length of search time, with all gains concentrated on low-skilled drivers. This result suggests that the AI is a substitute for worker skill, at least in our context. 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 findings are consistent with those of Webb (2020), in that skilled workers are more

[18] Table A5 verifies that our main results reported in Table 2 hold in this limited sample as well.
[19] Table A6 reports the results of using the full sample, which are very similar.

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, whereas Webb (2020) shows it *across* occupations. In addition, Webb (2020) argues 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 the productivity gap between high- and low-skilled workers.

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

This study faces several limitations. First, whereas we show that low-skilled drivers benefit from AI, one puzzle is that the utilization rate of AI is low even among the low-skilled drivers (5.9%).[20] One possibility is that the productivity gain of 10%, which translates into a reduction of search time by 1.5 minutes, is not large enough to make them recognize the improvement, especially because those low-skilled drivers may be inexperienced. Relatedly, over half of the drivers with the opportunity to use the application for free never bother to try it. In fact, 51.6% of these never-users are low-skilled drivers who would have benefited if they had used it.[21] While identifying the reasons for aversion to AI is beyond the scope of this study, drivers might feel competitive pressure from AI, according to which AI might replace their core skill of demand forecasting.[22] Second, we can speak little about the general equilibrium effect; what if all drivers in the area were to 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. Consumers are benefited, however, from the shorter time of finding a taxi. To the extent that this improved convenience stimulates further demand for a taxi, the market could expand and 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 a drastic transformation may

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

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

[22] This may not be a so-called algorithm aversion, that is, even when an algorithm consistently beats human judgment, people prefer to go with their own judgment (Dietvorst et al. 2015; 2016). Never-users, by definition, have not seen that the AI does better than their own judgment. In addition, information sharing among drivers seems rare according to our interviews with taxi drivers.

take time because the information required for driving tasks, such as the road environment, is much less regularized than information required for demand-forecasting tasks, such as passengers' location. As Autor (2015) points out, automating a task is much more costly under a non-regularized environment than a regularized environment, and the cost is likely to exceed the wage saving. As a result, the partial automation of tasks within an occupation (like this taxi case) is likely to persist in a wide range of occupations. 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 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 Press.

[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 Economics*, 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] Dhaene, G., and Jochmans, K. (2015) "Split-panel Jackknife Estimation of Fixed-effect Models," *Review of Economic Studies*, 82(3): 991–1030.

[17] 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.

[18] 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.

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

[20] 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.

[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] 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.

[24] Frey, C. B., and Osborne, M. A. (2014) "The future of employment," Working paper.

[25] Frey, C. B., and Osborne, M. A. (2017) "The Future of Employment: How Susceptible Are Jobs to Computerisation?," *Technological Forecasting and Social Change*, 114: 254–280.

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

[27] 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.

[28] 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.

[29] Imbens G. W. (2015) "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50(2): 373–419.

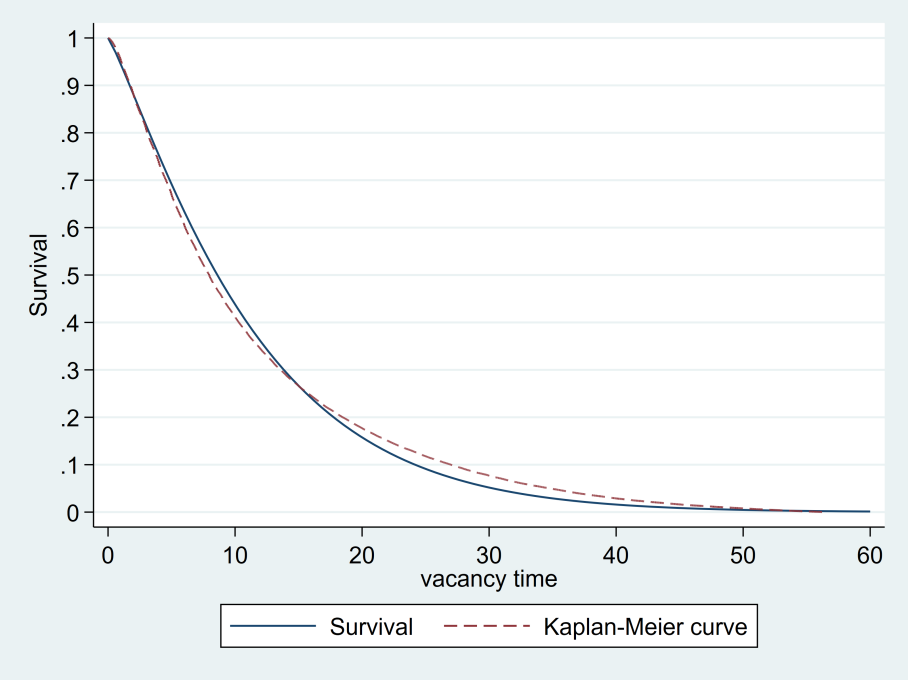
[30] Katz, L. F., and Murphy, K. M. (1992) "Changes in Relative Wages, 1963-1987: Supply and Demand Factors," *The Quarterly Journal of Economics*, 107(1): 35–78.

[31] Webb, M. (2020) "The Impact of Artificial Intelligence on the Labor Market," *SSRN working paper No. 3482150.*

[32] Wooldridge, J. M. (2010) *Econometric Analysis of Cross Section and Panel Data, 2nd ed.*, MIT Press.

Figures and tables

Figure 1: Model Prediction vs. Kaplan-Meier Curve



Notes: The figure 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 AI 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 taxis at the 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 at least once during the trial period. ***, ** and * denote 10%, 5%, and 1% significance levels, respectively.

Table 2: Weibull Hazard Regression: Main results

	(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 dummies for drivers whose skill index is above the median and below the 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 at least once during the trial period. Columns (3) and (6) further limit "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 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 levels, respectively.

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

	(1) Navi users -	(2) 4th week Navi users -	(3) Navi users low-skilled	(4) 4th week Navi users low-skilled	(5) Navi users high-skilled	(6) 4th week Navi users 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 × 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: The "Navi users sample" is used. High- and low-skilled are drivers whose skill index is above the median and below the median, respectively. "Navi usage" captures the impact of AI in the first two weeks, and "Navi usage × 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 at least once in the 4th week. Standard errors clustered on drivers are reported in parentheses. ***, ** and * denote 10%, 5%, and 1% significance levels, respectively.

Table 4: OLS Regression on Fare per Ride

	(1) Navi users	(2) Navi users	(3) Navi users	(4) Navi users	(5) Navi users $0.1 \leq PS \leq 0.9$	(6) Navi users $0.1 \leq PS \leq 0.9$
	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: The "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 dummies for drivers whose skill index is above the median and below the median, respectively. Columns (5) and (6) further limit 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 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 levels, respectively.

Online Appendix

(Not for Publication)

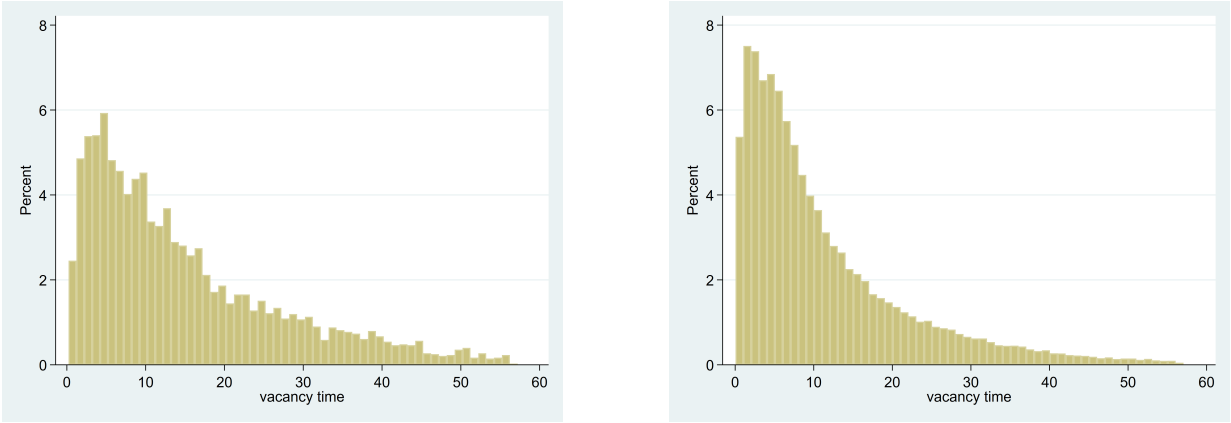
A. Additional figures and tables

Figure A1: Snapshot of AI Navi



Notes: The figure displays a 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

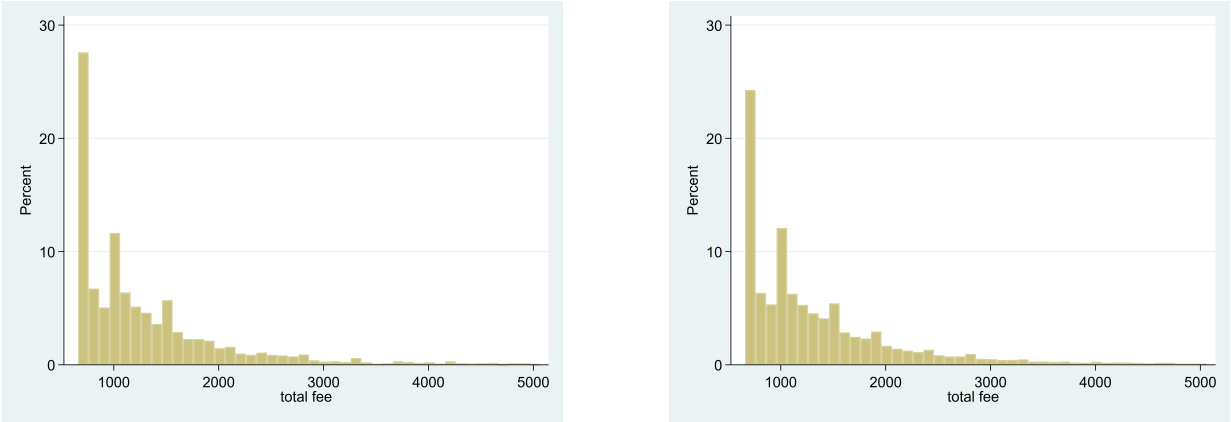


(a) AI is turned on

(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 it 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 it 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 the median and below the 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 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 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 levels, 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 the median and below the 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 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 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 levels, 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 quartiles by the distribution of skill index. "Skill index 1st quartile" is a dummy for drivers whose skill index is in the first quartile of the distribution. "Skill index 2nd quartile," "skill index 3rd quartile," and "skill index 4th quartile" are defined similarly. "Full" in column (1) is the sample of all drivers, and "Navi users" in column (2) is the sample of drivers who used AI at least once during the trial period. Column (3) further limits "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 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 levels, 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 AI's suggested routes that drivers did follow, calculated for each vacant cruise. Odd-numbered columns replicate the results of Table 2 for ease of comparison. Even-numbered columns add the interaction of AI usage dummy and Navi compliance rate to odd-numbered columns. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

Table A5: 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 the median and below the 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 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 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 levels, respectively.

Table A6: 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 the median and below the 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 levels, respectively.

B. Cox Proportional hazard model

Table B1 instead reports the estimates of the Cox Proportional hazard model to allow for a non-parametric baseline hazard. To make them comparable to the estimates of Weibull hazard model, Table B2 converts the estimates in Table 2 to hazard ratios. Note here that estimates greater than one mean that the probability of picking up a customer increases as the corresponding variable increases, while estimates less than one means that the probability decreases. Comparing Tables B1 and B2, the estimates from two models are almost identical.

Table B1: 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 more than one mean that the probability decreases, while estimates less than one mean that the probability decreases. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

Table B2: 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 ratios, so that they are comparable to the estimates from the Cox Proportional hazard regression reported in B1. Since the estimates here indicate hazard ratios, estimates greater than one mean that the probability of picking up a customer increases as the corresponding variable increases, while estimates less than one mean that the probability decreases. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

C. Incidental parameters problem

We address the incidental parameters problem due to estimating numerous fixed effects (FEs), specifically driver, date-hour, and ward FEs, in our non-linear model. The incidental parameters problem occurs because the number of observations within a fixed effect is small. To overcome this problem, we take three alternative approaches. Overall, we find that the bias on our estimates seems limited.

The first approach is to drop the fixed effects with a small number of observations, because the bias arising from these FEs are more severe. By dropping these observations, we can gauge the degree of bias arising from the incidental parameters problem. Specifically, we drop the observations involving the FEs with a number of observations less than the 10 percentile of the numbers of observations for driver, ward, and date-hour FEs.[23]

Table C1 reports the results of estimating our main model (1) with this trimmed sample. The estimates are very similar to the ones reported in Table 2. For instance, the estimate in column (1) of Table 2 is -0.051, whereas the current estimate is -0.046.

The second approach is to employ Mundlak (1978) to impose distributional assumptions on the fixed effects to reduce the number of parameters. More specifically, we assume that driver $\text{FE}_i = \gamma\overline{\text{Navi usage}}_{\text{i}} + c_i$, ward $\text{FE}_j = \delta\overline{\text{Navi usage}}_j + d_j$, and date-hour $\text{FE}_h = \theta\overline{\text{Navi usage}}_h + e_h$, where $\overline{\text{Navi usage}}_i$ is the mean of AI usage dummy for driver i, and $\overline{\text{Navi usage}}_j$ and $\overline{\text{Navi usage}}_h$ are analogously defined. The error terms c_i, d_j, and e_h are assumed to be strictly exogenous and treated as the random effects. The estimated parameters γ, δ, and θ indicate the direction and degree of selection.

Table C2 reports the results of estimating equation (1) where we replace driver, ward, and date-hour FEs by the average AI usage, namely $\overline{\text{Navi usage}}_{\text{i}}$, $\overline{\text{Navi usage}}_{\text{j}}$, and $\overline{\text{Navi usage}}_{\text{h}}$, respectively. Again, the estimates do not substantially differ from the ones reported in Table 2. The size of estimates becomes slightly larger than the main results when the sample is restricted to observations with a propensity score between 0.1 and 0.9. In addition, the estimates on the average AI usage indicate the absence of selection based on driver FE, but the presence of selection on date-hour and ward FEs. The positive estimates indicate that AI tends to be turned on during the times and places with longer vacant times.

The third approach is to draw on the bias correction using the split samples. Conceptually, the bias arising from the incidental parameters problem approaches zero when the number of observations used to estimate each fixed effect approaches infinity. Thus, when

[23] For example, for 522 drivers FE, we count the number of observations for each driver, rank them from the top to the bottom, and drop the observations of drivers in the lowest 10 percentile. We do the same for ward and date-hour FEs.

the panel data is equally split into two samples by the period, the bias of estimators using the split sample becomes larger than the bias of estimators using the full sample. Denote the estimator using the full sample as $\hat{\theta}$, the estimators using the first-half sample and second-half sample as $\hat{\theta}^1_{1/2}$ and $\hat{\theta}^2_{1/2}$. In the case of balanced panel data, Dhaene and Jochmans (2015) demonstrate that the estimator $2\hat{\theta} - \frac{1}{2}(\hat{\theta}^1_{1/2} + \hat{\theta}^2_{1/2})$ is an unbiased estimator because the biases cancel out. Our data are not a balanced panel and our model includes fixed effects other than driver fixed effects, and thus the bias correction formula of Dhaene and Jochmans (2015) does not strictly apply to our case. Regardless, we calculate the bias corrected estimator proposed by Dhaene and Jochmans (2015) as an approximation.

Tables C3 and C4 report estimates using the split samples by period. The comparison of the estimates using full sample reported in Table 2 (i.e., $\hat{\theta}$), the first-half sample reported in Table C3 (i.e., $\hat{\theta}^1_{1/2}$) and the second-half sample reported in Table C4 (i.e., $\hat{\theta}^2_{1/2}$) are close to each other. The similarity of three estimates suggests that the bias arising from the incidental parameters problem is limited. For instance, applying Dhaene and Jochmans's (2015) bias correction to column (1) of each table, the bias corrected estimate is -0.053[24], which is very close to -0.051, the baseline estimate reported in Table 2.

Overall, these exercises suggest that the bias due to the incidental parameters problem is fairly limited in our context. The primary reason is that the number of total observations (89,144) is large relative to the number of fixed effects.

References

[1] Dhaene, G., and Jochmans, K. (2015) "Split-panel Jackknife Estimation of Fixed-effect Models," *Review of Economic Studies*, 82(3): 991–1030.

[2] Mundlak, Y. (1978) "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46(1): 69–85.

[24] $= 2 \times (-0.051) - \frac{1}{2}(-0.047 - 0.051)$

Table C1: Weibull Hazard Regression: Dropping Fixed Effects Under 10% of Sample Size

	(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	-0.046**	-0.051***	-0.054**			
	(0.018)	(0.019)	(0.023)			
Navi usage×low-skilled				-0.094***	-0.110***	-0.093***
				(0.024)	(0.024)	(0.031)
Navi usage×high-skilled				-0.001	0.012	-0.009
				(0.025)	(0.027)	(0.035)
$\log(p)$	0.150***	0.140***	0.200***	0.150***	0.140***	0.200***
	(0.004)	(0.006)	(0.008)	(0.004)	(0.006)	(0.008)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	86,026	43,626	9,282	86,026	43,626	9,282
N of drivers	469	190	138	469	190	138
Log-likelihood	-121,175	-61,898	-12,700	-121,171	-61,891	-12,698

Notes: We dropped observations whose sample size of the fixed effects is less than 10% for each driver, ward, and date-hour FEs, respectively. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

Table C2: Weibull Hazard Regression: Average Navi Usage 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	-0.046***	-0.043**	-0.091***			
	(0.017)	(0.017)	(0.022)			
Navi usage×low-skilled				-0.086***	-0.078***	-0.119***
				(0.025)	(0.025)	(0.031)
Navi usage×high-skilled				0.018	0.012	-0.037
				(0.022)	(0.024)	(0.034)
$\overline{\text{Navi usage}}_{\text{i}}$	0.110	0.173	0.130	0.060	0.067	-0.042
	(0.101)	(0.108)	(0.173)	(0.093)	(0.100)	(0.158)
$\overline{\text{Navi usage}}_{\text{h}}$	2.276***	1.122***	1.112***	2.283***	1.139***	0.996***
	(0.173)	(0.134)	(0.277)	(0.157)	(0.122)	(0.250)
$\overline{\text{Navi usage}}_{\text{j}}$	2.880***	2.125***	2.875***	2.700***	1.811***	1.951**
	(0.749)	(0.523)	(0.779)	(0.587)	(0.444)	(0.791)
high-skilled				-0.248***	-0.213***	-0.223***
				(0.014)	(0.021)	(0.038)
$\log(p)$	0.105***	0.095***	0.105***	0.114***	0.101***	0.112***
	(0.004)	(0.005)	(0.009)	(0.004)	(0.005)	(0.009)
N	89,144	46,208	10,076	89,144	46,208	10,076
N of drivers	522	212	161	522	212	161
Log-likelihood	-129,479	-67,624.5	-14,623	-128,658	-67,333	-14,566

Notes: $\overline{\text{Navi usage}}_{\text{driver}}$ means average AI usage time ratio for all observations of the same driver. $\overline{\text{Navi usage}}_{\text{date-hour}}$ and $\overline{\text{Navi usage}}_{\text{ward}}$ are defined similarly. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

Table C3: Weibull Hazard Regression: 1st and 2nd Weeks only

	(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	-0.047**	-0.049**	-0.050*			
	(0.022)	(0.023)	(0.029)			
Navi usage×low-skilled				-0.095***	-0.097***	-0.100**
				(0.033)	(0.032)	(0.041)
Navi usage×high-skilled				-0.005	0.001	0.002
				(0.028)	(0.031)	(0.033)
$\log(p)$	0.151***	0.139***	0.186***	0.151***	0.139***	0.186***
	(0.005)	(0.006)	(0.010)	(0.005)	(0.006)	(0.010)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	39,495	20,826	6,544	39,495	20,826	6,544
N of drivers	511	212	161	511	212	161
Log-likelihood	-55,797	-29,636	-9,057	-55,795	-29,633	-9,055

Notes: Only the first two weeks of the sample are used. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.

Table C4: Weibull Hazard Regression: 3rd and 4th Weeks only

	(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**	-0.061**	-0.074**			
	(0.025)	(0.025)	(0.031)			
Navi usage×low-skilled				-0.089***	-0.110***	-0.079**
				(0.033)	(0.032)	(0.039)
Navi usage×high-skilled				-0.015	-0.005	-0.067
				(0.038)	(0.039)	(0.052)
$\log(p)$	0.157***	0.148***	0.244***	0.157***	0.148***	0.244***
	(0.005)	(0.007)	(0.013)	(0.005)	(0.007)	(0.013)
driver FE	✓	✓	✓	✓	✓	✓
ward FE	✓	✓	✓	✓	✓	✓
date-hour FE	✓	✓	✓	✓	✓	✓
N	49,649	25,382	3,532	49,649	25,382	3,532
N of drivers	501	209	88	501	209	88
Log-likelihood	-69,446	-35,766	-4,683	-69,445	-35,764	-4,683

Notes: Only the last two weeks of the sample are used. High- and low-skilled are a dummy for drivers whose skill index is above the median and below the 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 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 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 levels, respectively.