東京大学 公共政策大学院
ワーキング・ペーパーシリーズ
GraSPP Working Paper Series
The University of Tokyo

GraSPP-DP-E-22-001

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

Kyogo Kanazawa
Daiji Kawaguchi
Hitoshi Shigeoka
Yasutora Watanabe

October 2022

GraSPP
THE UNIVERSITY OF TOKYO
GraSPP Discussion Paper E-22-001

GRADUATE SCHOOL OF PUBLIC POLICY
THE UNIVERSITY OF TOKYO
HONGO, BUNKYO-KU, JAPAN

GraSPP
THE UNIVERSITY OF TOKYO

GraSPP-DP-E-22-001

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, IZA, and NBER**

October 2022

Graduate School of Public Policy
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, Japan
Phone:+81-3-5841-1349

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, IZA, and NBER

October 25, 2022

Abstract

We examine the impact of Artificial Intelligence (AI) on productivity in the context of taxi drivers. The AI we study assists drivers with finding customers by suggesting routes along which the demand is predicted to be high. We find that AI improves drivers' productivity by shortening the cruising time, and such gain is accrued only to low-skilled drivers, narrowing the productivity gap between high- and low-skilled drivers by 14%. The result indicates that AI's impact on human labor is more nuanced and complex than a job displacement story, which was the primary focus of existing studies.

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

*We thank Daisuke Adachi, Fernando Aragon, Krishna Pendakur, Kevin Schnepel, Kensuke Teshima, Shintaro Yamaguchi, Hongliang Zhang, and seminar participants at AASLE, Hitotsubashi University, Monash University, NBER Japan Project Meeting, Nihon University, Osaka University, and Simon Fraser 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-kyogo@g.ecc.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). The distributional consequences of AI could be fundamentally different from those of past technologies, such as IT and robotics, which are considered to be skill augmenting and inequality enhancing.[1] Whereas past technologies have replaced the routine and manual tasks of low-skilled workers, AI may replace non-routine cognitive tasks of typical high-skilled workers (Webb 2020).

Prior works on the labor-market consequences of AI focus primarily on job displacement across occupations–identifying the types of occupations that are more exposed and replaceable by AI.[2] To better understand the impacts of AI on the labor market, however, the fundamental factor to consider is how productivity, which plays a crucial role in determining employment and wages, is affected by AI. Furthermore, past studies implicitly assume that all workers within the occupation are uniformly affected by AI, missing substantial within-occupation heterogeneity of skills for the tasks that can be replaced by AI.[3]

To fill 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. First, a worker's individual productivity is easily measured by the length of time it takes to catch customers; each driver works independently and has considerable discretion as to how they find customers. Our data show that over one half of drivers' working time is devoted to searching for (and the remaining to carrying) customers, excluding breaks. Thus, customer search is among the most important tasks for taxi drivers. 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 utilize the same capital, charge the same prices, and face the same input prices.

The particular AI we study is called "AI Navi," which helps drivers find customers when a taxi is cruising. The AI suggests routes based on predicted demand to maximize the probability of catching customers, given the current location. Therefore, this AI is expected to improve the productivity of drivers by reducing search time. This type of AI, which increases

[1] E.g., Autor et al. (2003); Bartel et al. (2007); Acemoglu and Restrepo (2020, 2022).

[2] These studies are based on a task-based model of technology and labor (Acemoglu and Restrepo 2018), 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 highly exposed to AI (e.g., Felten et al. (2018, 2019); Frank et al. (2019); Webb (2020); Alekseeva et al. (2021).)

[3] 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 even more likely to do so. The primary difference is that we study the impact on productivity.

the accuracy of *prediction* tasks from the patterns of data using machine learning technology, is widely used in real business settings (Agrawal et al. 2018, 2019).[4] Demand forecasting–the process of predicting 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 on productivity 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 (FEs) to account for underlying demand, namely ward, and date-hour FEs. Our identifying assumption, thus, is that turning on the AI is quasi-random after controlling for these granular sets of FEs, and we empirically demonstrate the validity of this assumption. In addition, we use the drop-off location of the previous customer as an instrument for the driver's decision to turn on AI, assuming that the customer's destination choice is exogenous to the driver, and that the degree of (un)familiarity of the drop-off location is correlated with the driver's decision to turn on AI that assists in finding customers.

We find that AI improves the productivity of taxi drivers by shortening the search time by 5%, on average. Estimations with and without IV yields similar results, reassuring that the endogeneity of AI usage is not a serious concern, given the rich sets of fixed effects. Importantly, the productivity gain is concentrated on low-skilled drivers; the impact on low-skilled drivers, where skill is defined by previous driving performance, is 7%, whereas the impact on high-skilled drivers is nearly zero or even negative (albeit not statistically significant). As a result, the AI narrows the productivity gap between high- and low-skilled drivers by about 14%. These results indicate that the AI is a substitute for worker skill, at least in this context.

Our result has implications for the distributional consequences of AI. Most extant studies show that AI technologies predominately harm high-skilled occupations, as an AI substitutes for tasks that require the types of skills that high-skilled workers possess (See, e.g., Webb 2020). Instead, we show that, within an occupation, an AI can potentially benefit low-skilled workers while not affecting high-skilled workers much. While both results point out AI's potential to narrow the gap between high- and low-skilled workers, our result suggests a different channel through which the productivity gap may be reduced. Overall, our case indicates that the impact of AI on human labor is more nuanced and complex than a job replacement story.

[4] "The current generation of AI provides tools for prediction and little else." (Agrawal et al. 2018, p. 133).

Of course, this is a case study, and our findings speak only to the case of taxi drivers. Nonetheless, to the extent that the core skill of a job involves a prediction task from the patterns of data, and the AI improves the accuracy of this prediction task, our results may also be applicable to such occupations. For example, AI that reviews contracts for unusual clauses and AI that detects cells within malignant tumors might improve the productivity of low-skilled paralegals and low-skilled pathologists, respectively, more than those of their high-skilled counterparts.

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 3.77 million, the second largest in Japan after Tokyo. With an area of 435 km^2 (about 7 times the area of Manhattan), Yokohama city is divided into 18 wards. As of December 2019, there are 8,842 registered taxi drivers working for taxi-operating firms in the city. 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; JPY140 is approximately 1USD) and the variable charge after the first 2 km, which is determined by 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 and 60% based on our interviews with a random set of drivers), with a guaranteed baseline salary so that they will not work below minimum wage. Drivers do not incur any variable costs (including gas).

2.2 AI Navi

AI Navi, developed by a tech company, is designed to help drivers find customers when the taxi is cruising. Using a machine-learning technique, AI Navi's demand-forecasting capacity is based on recent driving records in Yokohama city.

More specifically, when turned on, AI Navi suggests routes for taxi drivers to maximize the probability that a taxi will catch customers, given the location of the taxi. Thus, AI is expected to improve the productivity of taxi drivers by reducing cruising time. Figure A1 displays a snapshot of AI Navi when it is turned on. AI Navi displays the suggested routes

in green with a red arrow, given a taxi's current location, and red dots indicate the locations with potential customers.

We prefer search time as a productivity measure over conventional outcomes (e.g., daily sales or hours worked), because existing studies on taxi drivers, such as Camerer et al. (1997), demonstrate that taxi drivers may *choose* the hours worked depending on their productivity (i.e., targeted income) and thus, such behavioral responses may contaminate the measurement of productivity. Also, as mentioned above, more than one half (57.96% from our data, excluding breaks) of a driver's working time is devoted to cruising, and hence improving the efficiency of search activities is critical for drivers. Finally, reducing search time is the direct objective of this particular AI.

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 roughly 500 taxi drivers ($\approx 6\%$) working for taxi-operating firms in Yokohama city *for free* during the period of December 3 to 31, 2019 (29 days). Taxi drivers who participated in this free trial received no reward or penalty for use or nonuse of the application. Therefore, it was entirely up to the discretion of taxi drivers whether 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 cruise during which drivers are searching for customers. Formally, we define a cruise as the time between when a cruise starts (i.e., dropping off the previous customer) and when it ends (i.e., picking up the next customer) on the street.

The original data consist of 67,111 cruises in December 2019. We make the following sample restrictions. First, we exclude the cruises of drivers whose pre-period data do not exist to construct our skill measures (N = 2,044). Second, following Haggag et al. (2017), cruises over 60 minutes are considered as being on a break and thus are excluded (N = 2,758). The final sample consists of 62,309 cruises of 520 drivers.

Among the 62,309 cruises, the number of cruises when AI is turned on and off is 3,127 (5.0%) and 59,182 (95.0%), respectively. Out of a total of 520 drivers, 201 (39%) used AI at least once and 319 (61%) never used it during the trial period. Thus, while 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. Generally, our results are robust to the use of either dataset.

Figure A2 shows the distributions of cruising time separately for (a) when AI is turned on, and (b) when AI is turned off. Both the mean and median cruising times are *higher* when AI is turned on than when it is turned off; the mean(median) time when AI is turned on is 15.6(11.4) minutes, whereas the time when it is turned off is 11.7(7.95) minutes, 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 cruising time between when AI is turned on and off is problematic, because it 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 cruising time when AI is turned on and off. As discussed earlier, however, we cannot simply compare the average cruising time between AI usage and non-usage, because the timing of AI usage could be endogenous. Thus, we compare the cruising 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 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 AI being turned on during a cruise. We assume that the duration of the cruise (in minutes), T, follows a Weibull distribution. The survival function, $S(t) = \Pr(T > t)$, which is the probability that drivers cannot find a customer until time t, by driver i in ward j at date hour h for cruise s 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 after 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 B; the bias is considered small. Parameter p captures the duration dependence of the baseline hazard, where $p = 1$ implies the absence of duration dependence, $p > 1$ ($log(p) > 0$) implies positive duration dependence, and $p < 1$ ($log(p) < 0$) implies negative duration dependence.

This model can be interpreted as

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

where ϵ follows an extreme-value distribution (Wooldridge 2010, p. 998). Our coefficient of interest is α, which corresponds to the percentage change in cruising 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 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) without a dummy of AI usage, regressing the cruising 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 a standard deviation of 1. Figure A3 shows the distribution of the skill index. This index essentially captures each driver's skill in finding 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 education and experience (e.g., Autor et al. 2008). Importantly, because this skill measure captures not only the demand-forecasting skill but also other skills necessary to shorten search time, such as driving skills,[5] it is not obvious *a priori* whether AI is more beneficial for more or less skilled drivers.

Similarly, we construct a vacancy index by estimating a hazard model similar to equation (1), regressing cruising 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 means more time to catch a customer, indicating a lower demand for a taxi at the ward-day-hour level,

[5]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.

on average.

3.2 Credibility of the underlying assumption

Recall that our identifying assumption is that turning on 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 their 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 (albeit not statistically significant), indicating that low-skilled drivers are more likely to use AI. 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. 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, whether to turn on or off AI is considered as good as random. This is plausible, because none of the drivers had been exposed to this application before, and thus they were likely to randomly experiment with it by turning it on and off.

Columns (3) and (4) in Table 1 repeat the same exercise for only those drivers who use AI at least once during the trial period ("Navi users" sample) and find similar patterns. Finally, column (5) adds driver FE to column (4), and the estimate on the vacancy index is again small.

3.3 Instrumental variable approach

The previous subsection shows that although AI is more likely to be used when demand is low, this underlying demand condition can be well-controlled by including a rich set of FEs. One may be still concerned, however, that unobserved demand conditions that are not fully controlled by FEs affect both the driver's decision to turn on AI and also the time to find the customer. To account for the remaining concern about the endogeneity of AI usage, we employ an instrumental variable (IV) approach. Specifically, we use two IVs: the drop-off location of the previous customer and the frequency of past AI usage.

The first IV is the drop-off location of the previous customer. The idea behind is that the

customer's destination choice is arguably exogenous to the driver, and that dropping off a customer (=starting a cruise) in an unfamiliar place induces the driver to turn on AI to assist their search. To measure the unfamiliarity, we calculate the share of cruises starting from each of 18 wards for each driver in the pre-trial period (October and November). Then, we call one minus this past share at each ward j for each driver i as the *unfamiliarity* index—the higher the more unfamiliar the location is for the driver. Then, we assign this index for any cruises s starting from ward j for driver i in December as an instrument.[6] To the extent that driver is more likely to turn on AI at unfamiliar locations, this IV can have predictive power of driver's decision to turn on AI (i.e., relevance).[7] Here, if starting from unfamiliar places directly affects the length of search time regardless of AI usage, the IV violates the exclusion restriction. To address this, we always control for the average cruising time for each driver i at ward j using the pre-trial period (called the Cruising time index, hereafter)[8] in this IV approach, which turns out not to be empirically important. The second IV is simply the number of AI usages until the current cruise s for each driver i. The idea is that drivers should accumulate experience of AI usage by past quasi-random events, and thus past AI usage induces current AI usage.

Recall that our model of interest is a survival model with a time-varying binary treatment variable that considers the hazard of finding a customer, given cruising time. Thus, we need the predicted probability of turning on AI that varies with cruising time. To approximate this time-varying probability, we estimate the Tobit model where the dependent variable is time to turn on AI where the upper bound is the end of cruise.[9] We then transform the predicted time of turning on AI from the Tobit into the time-varying probability of turning

[6]For simplicity, suppose that there are only three wards and that driver i's share of the cruises starting from Wards 1, 2, and 3 in October and November are 0.7, 0.2, and 0.1, respectively. Then, for any cruises s starting from Wards 1, 2, and 3 in December for driver i, we assign 0.3, 0.8, and 0.9 (= 1-past share), respectively. The relevance of such an IV means that driver i is more likely to turn on AI when the cruise starts in Ward 3 (= 0.9) than in Ward 1 (= 0.3).

[7]One concern is that the most drivers pick up and drop off within very limited locations, and thus we may lack statistical power. However, Figure A4 plots the distribution of the share of cruises stating from the Naka Ward (the share is the highest at 37.1%) at the driver level and reveals substantial variation across drivers. Consequently, the Herfindahl-Hirschman index (HHI) of the unfamiliarity index in Figure A5 is widely dispersed.

[8]More precisely, we regress cruising time on driver-ward FE and date-hour FE using data from October and November and use the estimated driver-ward FE as the Cruising time index.

[9]To capture the probability of turning on AI at any given time, one may think that a competing hazard model with two terminal points, turning on AI and finding customers, may be a proper model. We do not take this avenue, however, because strictly speaking, our case does not fit the standard competing hazard model, in that a cruise continues even after AI is turned on; thus an end-point event (= finding a customer) follows after another end-point event (= turning on AI). Moreover, estimating a competing hazard model with many fixed effects is computationally not feasible.

on AI.[10]

Table 2 shows the results of our first-stage Tobit regression where we regress the log of time until AI is turned on on two IVs in columns (1) and (2), and further their interaction in columns (3) and (4). As expected, both IVs are negative and statistically significant, indicating that drivers turn on AI faster at unfamiliar locations, and as they are more experienced with AI Navi. To be sure, the shorter time to turn on AI implies a higher probability of turning on AI at any given time.

The practical challenge for our IV approach is that, to our knowledge, there is no standard way of adopting IVs in a non-linear model like the Weibull hazard model. Thus, we take a rather ad hoc approach and add the residual from Table 2 to equation (1), in the spirit of the control function approach. Here, we assume that the residual of the Tobit model (i.e., the variation of time to turn on AI *not* explained by IV) captures the endogeneous variation of AI usage.[11] It turns out that the specifications with and without this residual control yield similar results (as shown below), implying that the remaining endogenity does not appear to be a major concern.

4 Results

4.1 Overall productivity

Table 3 reports the main result of estimating equation (1). Column (1) shows that AI reduces the time spent on cruising by 5.1% using the full sample. We show graphically the fitness of our hazard model. Figure A6 compares the estimated survival curve of column (1) of Table 3 (solid) and the Kaplan-Meier curve (dash). The extent of the fit is reasonably high, suggesting that the Weibull distribution captures the underlying hazard well. Column (2) limits the sample to AI Navi users and finds similar results.[12] To account for the low utilization rate of AI, we trim the sample based on the propensity score in column (3), to ensure sufficient overlap in characteristics between cruises with and without AI usage. Specifically, we calculate the propensity score to turn on AI from the logistic regression of the AI usage

[10] First, the time at which AI is turned on for each cruise is predicted from the Tobit model in Table 2. Second, the "AI turn-on probability" at the predicted timing in the first step is set to 0.5 and the probability at the start of the cruise is set to 0. Third, each cruise is divided into 1-minute segments, and "AI turning-on probability" is linearly interpolated and extrapolated for each segment of each cruise.

[11] The difference between the "AI turning-on probability" and the (actual) AI usage dummy becomes the residual. We always use this 1-minute split sample in the IV approach.

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

dummy on driver, ward, and date-hour FEs. Column (3) limits the sample in column (2) to cruises whose propensity score lies between 0.1 and 0.9 (Imbens 2015).[13] 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 positive estimates of $\log(p)$ in all specifications imply positive duration dependence, reflecting that drivers tend to move to locations where catching customers is easier.

Column (4) in Table 3 reports the results of the IV approach using the full sample where we control for the residuals calculated from the predicted time-varying AI turning-on probability from column (3) of Table 2. We are reassured that the IV approach yields similar estimates as those in column (1) without such residual controls, implying that role of remaining endogeneity is limited. Column (5) reports the estimate of IV approach using the Navi user sample, and again yields similar estimates as column (2) without residual controls.

4.2 Productivity gain by skills

We now report the productivity improvement by skill level. Figure 1 plots the estimated value of AI Navi effects by skill where we add the interactions of AI usage and the skill index and the square of skill index to equation (1). The figure shows that the estimate of AI usage on cruising time is constantly negative for a wide range of the low skill index, suggesting that productivity gains are concentrated on low-skilled drivers. In contrast, the estimates on high-killed drivers are not statistically distinguishable from zero, and possibly positive. Figure A7 reports similar results when the interaction of AI usage and cubic is added to the equation (1), suggesting that our results are not driven by the particular parametric form of the skill index.[14]

To quantify the magnitudes of productivity improvement by skill level, Table 4 reports the estimates from the specification, where the skill index that is divided by median, tertile, and quartile, respectively, is interacted with the AI usage dummy. Consistent with Figure 1, productivity gains are concentrated on low-skilled drivers. For example, column (2) of Table 4 shows that while AI reduces cruising time by 7.4% and 6.1% for the low- and middle-skilled tertiles, the corresponding gain for the high-skilled tertile is essentially zero (with large SE), suggesting that productivity gains are mostly accrued to drivers at the bottom two-thirds

[13] 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).

[14] See Table A1 for the corresponding estimates of square and cubic specifications.

of the skill index. As a result, the AI narrows the productivity gap between high- and low-skilled tertiles by 14%.[15] Nonetheless, the 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. Again, we are reassured that columns (4)-(6) with the IV approach yield similar results to columns (1)-(3) without residual controls.

5 Supplementary analysis

Compliance.— One remaining concern could be that even though the AI would also have benefited high-skilled drivers, these drivers simply did not follow the navigation routes suggested by the AI. Since AI Navi assists with 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 control for the "Navi compliance rate," which is the fraction of the AI's suggested routes that drivers did follow, calculated for each cruise.[17] Table A2 presents the results. Odd-numbered columns replicate Table 4 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 ("AI usage $\times$ skill index") is hardly changed. Therefore, it is unlikely that our results are driven by compliance.

The impact of AI over time.— We also investigate whether the impact of AI evolves over time. We split the sample period into the first two weeks and the second two weeks. Table A3 shows that AI's positive impact is immediate and observed already in the first two weeks, which 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. We do not see any improvement in the last two weeks (labeled "AI usage $\times$ 3rd/4th weeks"), suggesting that learning is limited in this setting, probably due to the short duration of the trial period.

[15] The difference of estimates between low- and high-skilled tertiles (-0.074-(-0.002)) from column (2) of Table 4) is divided by the difference of the average driver FE for low- and high-skilled tertiles (0.661-0.151)).

[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] In fact, the compliance rate for the high-skilled half and the low-skilled half are very similar (64.1% and 62.1%, respectively)). This rate is calculated by the tech company based on their definition.

Other model.—Appendix Section C reports the estimates of the Cox Proportional hazard model to allow for a non-parametric baseline hazard. We are reassured that the estimates from this model are almost identical to those of our baseline Weibull hazard model, suggesting that our results are not driven by the particular choice of hazard.

Another productivity measure.—Thus far, we use the length of search time as a measure of productivity, but another natural candidate would be 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 find customers with potentially long rides. Appendix Section D reports the results of the OLS regression on fare per ride. Overall, we do not find an economically large or statistically significant effect overall, nor by skill level, suggesting that AI does not seem to direct (low-skilled) drivers to locations with low-hanging fruit.

6 Discussion and Conclusion

We investigate the impact of AI on worker productivity in the context of taxi drivers. We find that AI improves productivity, measured by the length of search time, with all gains concentrated on low-skilled drivers. To the extent that productivity is reflected by wages (which we do not observe in our data), AI has the potential to reduce wage inequality across workers within the same occupation.

This study faces several limitations. First, while we show that low-skilled drivers benefit from AI, one puzzle is that the utilization rate of AI is low even among low-skilled drivers (5.2%).[18] One possibility is that the productivity gain of 7%—which translates into a reduction of search time by 0.96 minutes—is not large enough to make drivers recognize the improvement, especially because those low-skilled drivers may be inexperienced. Relatedly, more than one half of drivers who have the opportunity to use the application for free never bother to try it. In fact, 48.6% of these never-users are low-skilled drivers who would have benefited if they had used it.[19] While identifying the reasons for aversion to AI is beyond the scope of this study, drivers may simply be reluctant to adopt new technology. Second, we can speak little about the general equilibrium effect: What if all drivers in the area adopt this AI technology? One concern is that taxi drivers in the area would compete for the

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

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

same customers and end up engaging in business stealing if the market size stays constant. Consumers benefit, however, from the shorter time to find a taxi. To the extent that this improved convenience stimulates further demand for taxis, 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.[20] Agarwal et al. (2019) classify the type of AI usage in our setting as "augmenting labor on decision tasks," where the automation of prediction through AI can improve human decision making and consequently the productivity of labor. To the extent that the core skill of jobs involves a prediction task from patterns of data, such as paralegals to identify unusual clauses and pathologists to detect malign tumors, our results may also be applicable to such occupations. Whether our finding can be generalized to other settings is left for future research.

[20] 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 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 costlier under a non-regularized environment than a regularized environment, and the cost is likely to exceed the wage saving.

References

[1] Alekseeva, L, Azar, J., Gine, M., Samila, S., and Taska, B. (2021) "The demand for AI skills in the labor market," *Labour Economics*, 71, 102002.

[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, and Restrepo, P. (2022) "Tasks, Automation, and the Rise in U.S. Wage Inequality.," *Econometrica*, 90(5): 1973–2016.

[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] 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.

[13] 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.

[14] 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.

[15] Felten, E. W., Raj, M., and Seamans, R. (2019) "The Occupational Impact of Artificial Intelligence: Labor, Skills, and Polarization," *SSRN working paper No. 3368605.*

[16] 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.

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

[18] 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.

[19] 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.

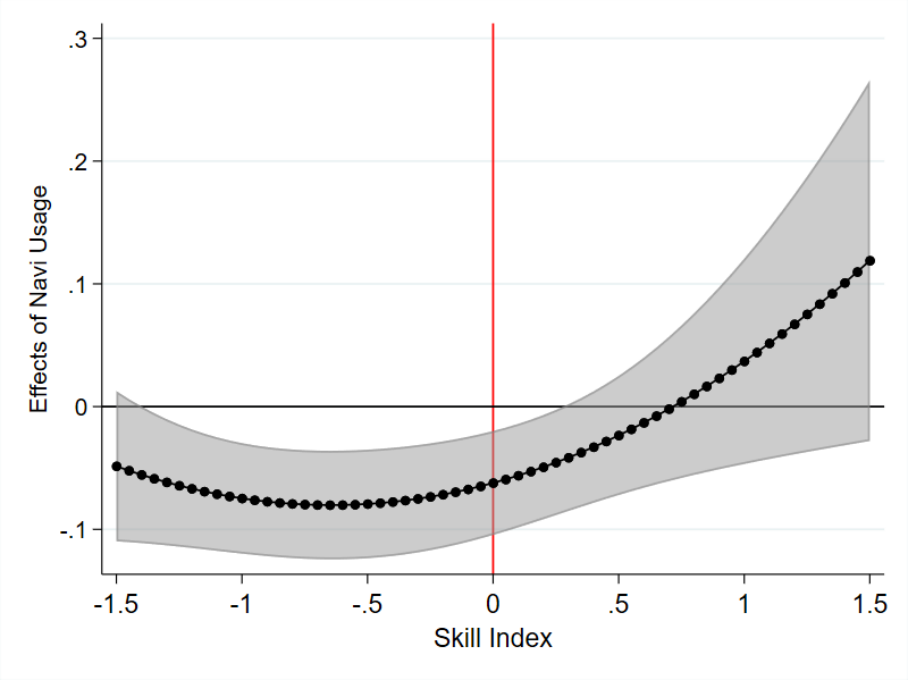
[20] Imbens G. W. (2015) "Matching Methods in Practice: Three Examples," *Journal of Human Resources*, 50(2): 373–419.

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

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

Figures and tables

Figure 1: AI effects by skill



Note: This figure plots the estimated value of AI Navi effects by skill with 90% confidence intervals in the shaded areas. The Weibull hazard regression, where the interactions of AI usage dummy and skill index and square of skill index are added to equation (1), is estimated. See column (1) of Table A1 for the corresponding estimates. The outcome is cruising time. The negative estimate indicates that AI usage reduces the search time. Higher skill index indicates more skilled drivers. Figure A3 displays the distribution of the skill index. Drivers whose skill index ranges from -1.5 to 1.5 cover 92.4% of all cruises. The "full" sample with all drivers is used.

Table 1: Determinants of AI usage

	(1)	(2)	(3)	(4)	(5)
Outcome:	AI usage dummy				
Sample:	Full	Full	Navi users	Navi users	Navi users
Skill index	-0.194	-0.183	-0.190	-0.172	
	(0.138)	(0.144)	(0.116)	(0.118)	
Vacancy index	0.110***	-0.037	0.102***	-0.027	-0.033
	(0.034)	(0.032)	(0.035)	(0.035)	(0.056)
Driver FE					✓
Ward FE		✓		✓	✓
Date-hour FE		✓		✓	✓
N	62,182	55,408	28,310	25,415	25,259
N of drivers	520	520	201	201	199
Log-likelihood	-12,338	-10,971	-9,773	-8,455	-5,066

Note: Estimates from the logistic regression are reported. The outcome is AI usage, which is a dummy that takes one when AI Navi is turned on. "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. Higher skill index indicates more skilled drivers. Table A3 displays the distribution of skill index. Higher vacancy index indicates less demand for taxis at the ward-day-hour level. Sample sizes are not identical to those of other tables due to observations with missing vacancy index and observations that are dropped by including fixed effects. Standard errors clustered on drivers are reported in parentheses. ***, **, and * denote 10%, 5%, and 1% significance levels, respectively.

Table 2: First stage—AI usage

	(1)	(2)	(3)	(4)
Outcome:	ln(Time until AI is Turned On)			
Sample:	Full	Navi users	Full	Navi users
Unfamiliar index	-0.541*	-0.550*	-0.761**	-0.800**
(= 1 - Past share of starting ward)	(0.324)	(0.333)	(0.354)	(0.363)
# of AI usage	-0.028***	-0.029***	-0.035***	-0.037***
	(0.004)	(0.004)	(0.007)	(0.007)
Unfamiliar index × # of AI usage			0.010	0.011
			(0.008)	(0.008)
Cruising time index	0.008	0.011	0.008	0.011
	(0.011)	(0.011)	(0.011)	(0.011)
Driver FE	✓	✓	✓	✓
Ward FE	✓	✓	✓	✓
Date-hour FE	✓	✓	✓	✓
N	62,309	28,369	62,309	28,369
N of drivers	520	201	520	201
Log-likelihood	-10,916	-10,688	-10,915	-10,687
F-stat	F(2, 61783)	F(2, 28159)	F(3, 61789)	F(3, 28162)
	= 21.71	= 24.61	= 14.23	= 16.01
p-value	<0.0001	<0.0001	<0.0001	<0.0001

Note: Estimates from the Tobit regression are reported. The outcome is logged time until AI is turned on. The unfamiliar index is one minus the "past share of ending ward," which is the past share of the cruises that started at each of 18 wards in the pre-trial period (October and November 2019) for each driver. "# of AI usage" is the number of AI usages till the current cruise for each driver. The "Cruising time index" is the average cruising time at each ward for each driver using data from the pre-trial period to control for the mechanical correlation between unfamiliar location and longer cruising time. "Full" in columns (1) and (3) is the sample of all drivers, and "Navi users" in columns (2) and (4) is the sample of drivers who used AI at least once during the trial period. Standard errors clustered on drivers are reported in parentheses. ***, **, and * denote 10%, 5%, and 1% significance levels, respectively.

Table 3: Overall AI effect

	(1)	(2)	(3)	(4)	(5)
Sample:	Full	Navi users	Navi users	Full	Navi users
Specification:	-	-	-	IV	IV
			$0.1 \leq PS \leq 0.9$		
AI usage	-0.051**	-0.046**	-0.067**	-0.047**	-0.040*
	(0.022)	(0.023)	(0.027)	(0.021)	(0.022)
Residual				-0.005***	-0.005***
				(0.001)	(0.001)
Cruising time index				0.021***	0.020***
				(0.001)	(0.002)
log(p)	0.179***	0.186***	0.247***	0.183***	0.191***
	(0.005)	(0.007)	(0.011)	(0.005)	(0.007)
Driver FE	✓	✓	✓	✓	✓
Ward FE	✓	✓	✓	✓	✓
Date-hour FE	✓	✓	✓	✓	✓
N	62,309	28,369	6,412	62,309	28,369
N of drivers	520	201	173	520	201
Log-likelihood	-85,505	-38,776	-8,464	-85,282	-38,681

Note: Estimates from the Weibull hazard regression of equation (1) are reported. The outcome is cruising time. "Full" in columns (1) and (4) is the sample of all drivers, and "Navi users" in columns (2), (3), and (5) is the sample of drivers who used AI at least once during the trial period. Column (3) further limits the sample in column (2) to 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. Column (4) controls for residuals calculated from the predicted AI turning-on probability from column (3) of Table 2, and column (5) controls for the residual from column (4) of Table 2. The "Cruising time index," which is the average cruising time at each ward for each driver using the pre-trial period data, is also controlled. Standard errors clustered on drivers are reported in parentheses. ***, **, and * denote 10%, 5%, and 1% significance levels, respectively.

Table 4: Heterogeneous AI effects by skill

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	Full	Full	Full	Full	Full	Full
Specification	-	-	-	IV	IV	IV
AI usage×Low-skilled half	-0.072***			-0.067***		
	(0.026)			(0.025)		
AI usage×High-skilled half	-0.025			-0.023		
	(0.035)			(0.034)		
AI usage×Low-skilled tertile		-0.074**			-0.067**	
		(0.030)			(0.029)	
AI usage×Middle-skilled tertile		-0.061			-0.062	
		(0.040)			(0.038)	
AI usage×High-skilled tertile		-0.002			0.006	
		(0.040)			(0.040)	
AI usage×Lowest-skilled quartile			-0.070*			-0.065*
			(0.038)			(0.037)
AI usage×Low-skilled quartile			-0.074**			-0.069**
			(0.036)			(0.035)
AI usage×High-skilled quartile			-0.048			-0.050
			(0.042)			(0.041)
AI usage×Highest-skilled quartile			0.035			0.047
			(0.059)			(0.057)
Residual				-0.005***	-0.005***	-0.005***
				(0.001)	(0.001)	(0.001)
Cruising time index				0.021***	0.021***	0.021***
				(0.001)	(0.001)	(0.001)
$\log(p)$	0.179***	0.179***	0.179***	0.183***	0.183***	0.183***
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
Driver FE	✓	✓	✓	✓	✓	✓
Ward FE	✓	✓	✓	✓	✓	✓
Date-hour FE	✓	✓	✓	✓	✓	✓
N	62,309	62,309	62,309	62,309	62,309	62,309
N of drivers	520	520	520	520	520	520
Log-likelihood	-85,504	-85,504	-85,503	-85,281	-85,281	-85,280

Note: Estimates from the Weibull hazard regression, where the interaction of AI usage dummy and skill index is included to equation (1) instead of the AI usage dummy alone, are reported. The outcome is cruising time. Low- and high-skilled halves in columns (1) and (4) are dummies for drivers whose skill index is below the median and above the median, respectively. Low-, middle-, and high-skilled tertiles in columns (2) and (5) are dummies for drivers whose skill index is below the first tertile, between the first tertile and second tertile, and above the second tertile, respectively. The skill index dummies for each quartile in columns (3) and (6) are similarly constructed. Columns (4)-(6) control for residuals calculated from the predicted AI turning-on probability from column (3) of Table 2. The "Cruising time index," which is the average cruising time at each ward for each driver using the pre-trial period data, is also controlled. The "full" sample with all drivers is used. Standard errors clustered on drivers are reported in parentheses. ***, **, and * denote 10%, 5%, and 1% significance levels, respectively.

Online Appendix

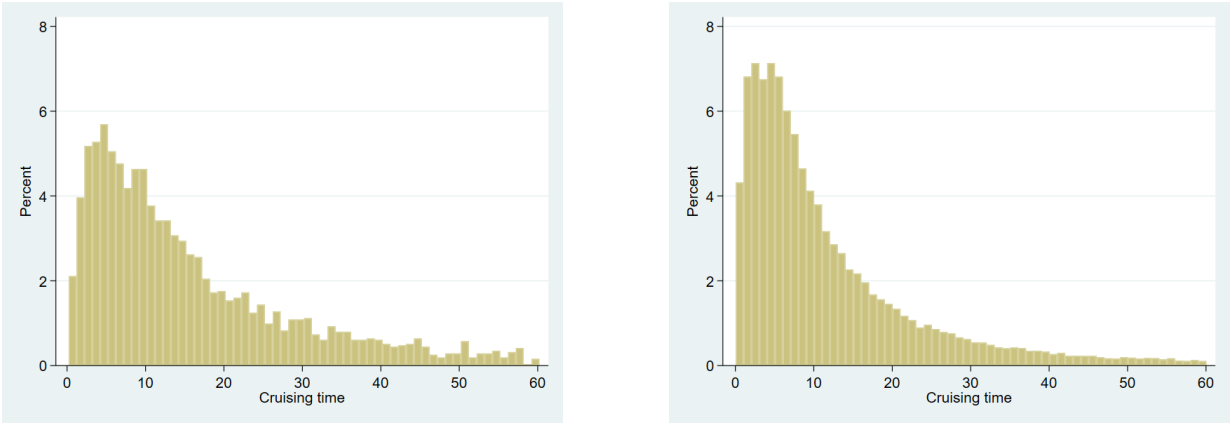
(Not for Publication)

A. Additional figures and tables

Figure A1: Snapshot of AI Navi

Note: 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.

Figure A2: Histogram of cruising time when AI is turned on/off

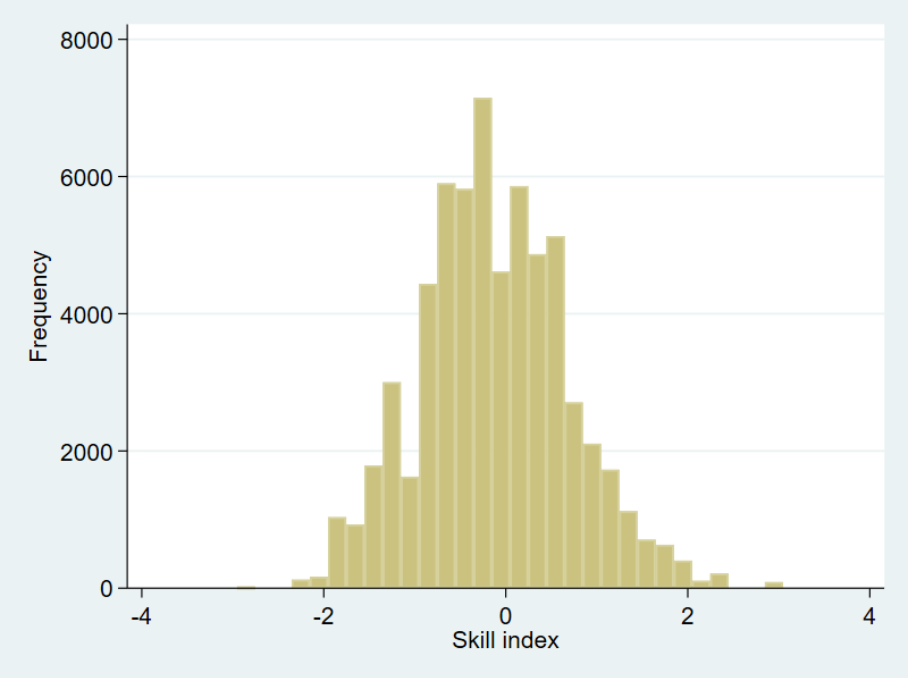


(a) AI is turned on

(b) AI is turned off

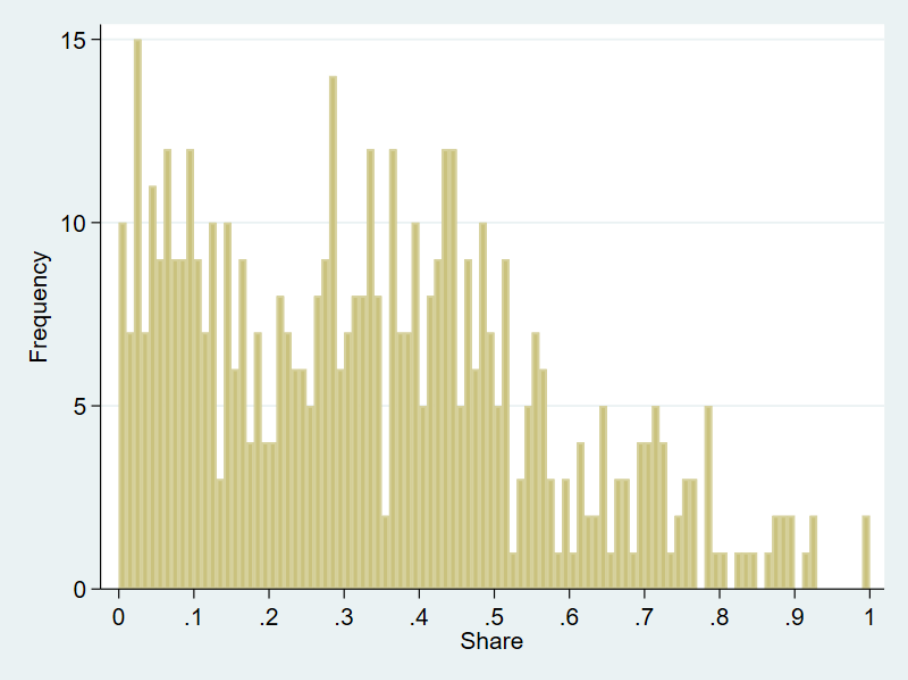
Note: These figures show the distributions of cruising time in the sample period (a) when AI is turned on, and (b) when AI is turned off separately. The mean(median) time (a) when AI is turned on is 15.6(11.4) minutes, whereas (b) when AI is turned off it is 11.7(7.95) minutes. The "full" sample with all drivers is used. The number of observations for (a) and (b) are 3,127 and 59,182, respectively.

Figure A3: Histogram of the skill index



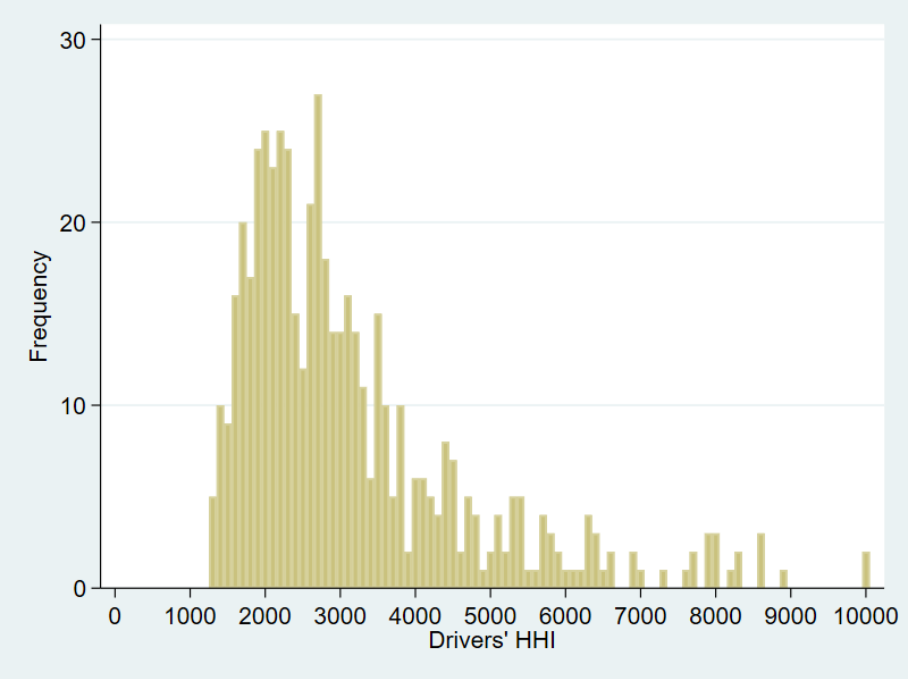
Note: This figure plots the distribution of driver's skill index at the cruise level. The skill index is constructed as follows. First, we estimate the Weibull hazard model of equation (1) without an AI usage dummy, regressing the cruising time onto driver, ward, and date-hour FEs. Then, we flip the 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 a standard deviation of 1.

Figure A4: Histogram of the share of cruises starting from Naka Ward



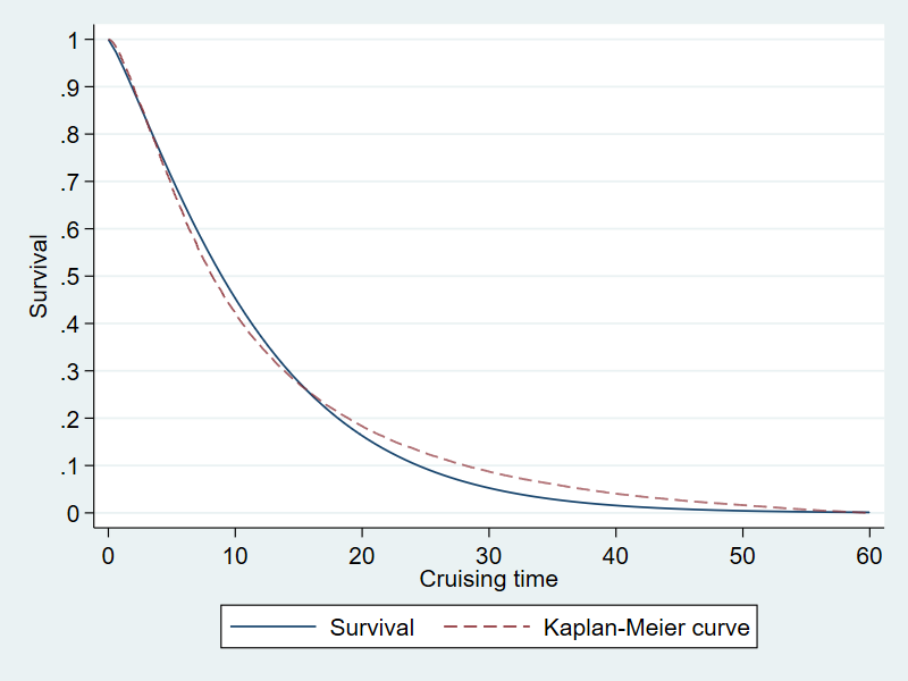
Note: The figure plots the distribution of the share of cruises that starts from Naka Ward, using data from the pre-trial period (October and November 2019). The unit of observation is driver (N= 520). The value of one indicates that all the cruises in the period for the driver start from Naka Ward, while the value of zero indicates that none of the cruises in the period for the driver start from Naka Ward. Among 18 wards in Yokohama-city, Naka Ward has the largest numbers of cruises that start from there (the share of 37.1%).

Figure A5: Histogram of the HHI of the shares of starting wards



Note: This figure plots the distribution of the Herfindahl-Hirschman Index (HHI) of the past share of the cruises starting at each ward. The unit of observation is driver (N= 520). Higher index indicates higher concentration of the cruises that starts from particular wards.

Figure A6: Model Prediction vs. Kaplan-Meier Curve



Note: The figure compares the estimated survival curve from column (1) of Table 3 (solid) from the Weibull hazard regression of equation (1), and the Kaplan-Meier curve (dash).

Figure A7: Heterogeneous AI effects by skill (cubic specification)

Note: This figure plots the estimated value of AI Navi effects by skill with 90% confidence intervals in the shaded areas. The Weibull hazard regression, where the interactions of AI usage dummy and skill index, square of skill index, and cubic of skill index are added to equation (1), is estimated. See column (2) of Table A1 for corresponding estimates. The outcome is cruising time. A negative estimate indicates that AI usage reduces the search time. A higher skill index indicates more skilled drivers. Table A3 displays the distribution of skill index. The drivers whose skill index ranges from -1.5 and 1.5 cover 92.4% of all cruises. The "full" sample with all drivers is used.