Behavioral measures improve AI hiring: A field experiment[1]

Marie-Pierre Dargnies (Univ. Lille, CNRS, IESEG School of Management, UMR 9221–LEM–Lille Économie Management, F-59000 Lille, France)

Rustamdjan Hakimov (University of Lausanne)

Dorothea Kübler (WZB Berlin, Technische Universität Berlin & CESIfo)

July 2026

Abstract

The adoption of Artificial Intelligence (AI) for hiring processes is often impeded by a scarcity of comprehensive employee data. We hypothesize that the inclusion of behavioral measures elicited from applicants can enhance the predictive accuracy of AI in hiring. We study this hypothesis in the context of microfinance loan officers, running an RCT where AI makes hiring decisions for half of the applicants. Our findings suggest that survey-based behavioral measures markedly improve the predictions of a random-forest algorithm trained to predict productivity within sample relative to demographic information alone. We then validate the algorithm's robustness to the selectivity of the training sample and potential strategic responses by applicants by running two out-of-sample tests: one forecasting the future performance of novice employees, and another with a field experiment on hiring. Both tests corroborate the effectiveness of incorporating behavioral data to predict performance and the usefulness of AI hiring. We find no evidence that applicants successfully manipulate the algorithm, which may reflect the difficulty of inferring optimal responses given the non-monotonic mapping from inputs to predictions.

Keywords: Hiring; AI; personnel economics; economic and behavioral measures; selective labels; strategic response to AI

JEL codes: D23, D91, M51

[1]We thank Brice Corgnet, Guido Friebel, Johannes Johnen, Amma Panin, Christian Zehnder, and participants of the 2025 European Decision Sciences Research Day at INSEAD (Fontainebleau), the 2023 Zurich Workshop on Economics and Psychology, the CREST workshop on experimental economics 2023, the Heilbronn Workshop on Field Experiments in Economics and Business 2024, and the Newcastle Experimental Economics Workshop 2024, COPE 2026, seminar participants at University of Lille, Málaga, Bamberg, Lingnan University Hong Kong, Lund University, NUS Singapore, Utrecht University, ECARES Brussels, University of Southern Denmark, University of Milan Biccoca, GATE Lyon, NES, HSE St. Petersburg, UC Louvain, University of Düsseldorf, and MPI Bonn. Marie-Pierre Dargnies acknowledges financial support by the ANR (ANR-20-CE26-0005 TrustSciTruths). Rustamdjan Hakimov acknowledges financial support by the Swiss National Science Foundation (project 100018_232179). Dorothea Kübler acknowledges financial support by the Deutsche Forschungsgemeinschaft through CRC TRR190 ("Rationality and Competition"). The study was pre-registered at the AEA RCT Registry (AEARCTR-0008219).

1 Introduction

Artificial intelligence, and machine learning in particular, has progressed rapidly in recent years (Goodfellow et al. 2016, Bengio et al. 2023). While earlier automation primarily affected jobs with a high share of routine tasks (Autor 2015, Arntz et al. 2016), machine-learning systems are increasingly applied to screening and prediction problems that traditionally relied on human judgment. Across several domains, algorithmic predictions can match or outperform expert decision-making, including in contexts related to selection and personnel decisions (e.g., McKinney et al. 2020, Mullainathan and Obermeyer 2022, Kleinberg et al. 2018, Ash et al. 2020, Chalfin et al. 2016). In labor markets, the diffusion of electronic applications has expanded applicant pools and increased the demand for scalable screening, contributing to the rapid spread of vendor-provided tools that score CVs and asynchronous interviews.[2] A central concern is fit: these largely 'one-size-fits-all' tools are typically developed and trained outside the client firm, so their effectiveness and fairness need not transfer across firms, jobs, and local labor markets. Consistent with this concern, audits document limited external stability—small changes in the format of input to measure applicants' personality (e.g., CV versus LinkedIn) lead to changes in AI output (Rhea et al. 2022 and references therein).

Developing and deploying a firm-specific hiring algorithm is difficult for several reasons. First, many firms—especially smaller ones—lack large, labeled employee datasets and the internal expertise needed to train and maintain predictive models (Agrawal et al. 2019, Radhakrishnan et al. 2020, Bhalerao et al. 2022). Second, even if a firm can train a model, the training sample is selective: performance outcomes are observed only for candidates who were hired, and remained long enough for performance to be reliably measured, creating a "selective labels" problem that can distort predictions at the hiring margin (Chalfin et al. 2016; Kleinberg et al. 2018). How large this distortion is in typical firm settings is rarely known ex ante. Third, hiring creates incentives to manipulate inputs: when applicants understand that survey responses or other features are used for automated scoring, they may tailor answers strategically, and the extent to which such

[2] A vast majority of firms is already using AI tools for screening applicants, see https://www.forbes.com/sites/janehanson/2023/09/30/ai-is-replacing-humans-in-the-interview-processwhat-you-need-to-know-to-crush-your-next-video-interview/ (last accessed on 9.11.2025).
The market for such tools is predicted to expand further, https://www.maximizemarketresearch.com/market-report/global-ai-recruitment-market/63261/ (last accessed on 20.02.2025), with new methods emerging continuously (e.g., Guenzel et al. 2026).

behavior degrades predictive performance remains unclear.[3] This paper asks whether firms with thin administrative data can nonetheless develop effective, firm-specific hiring algorithms by augmenting records with standardized behavioral measures, and how much selective labels and strategic responding matter at deployment. We quantify the value of algorithmic hiring in a randomized field experiment that compares algorithmic hiring to the firm's status-quo practice, and we benchmark the incremental value of survey measures on top of administrative data.

We collaborate with a microfinance company in Kyrgyzstan. In the microfinance sector, hiring prerequisites are minimal, and the firm collects only a small set of employee characteristics at entry. At the same time, productivity varies substantially across loan officers, and turnover is concentrated among low performers. These features make the setting particularly relevant for studying how a firm can build a hiring tool when administrative data are thin but performance differences are economically large.

To address data scarcity, all loan officers were required to complete a survey that elicits standardized measures developed and widely used in economics and psychology. The survey includes both incentivized and non-incentivized economic measures, eliciting risk and time preferences, trust, trustworthiness, and altruism, alongside psychological measures and cognitive tests. These measures of behavioral traits and preferences correlate with education, employment, and related outcomes (Barsky et al. 1997, Dohmen et al. 2011, Gottfredson 2002, Dohmen et al. 2009, Buser et al. 2014, Hakimov et al. 2023, Alan et al. 2019, Hanushek et al. 2023). Personality testing is also common in practice, though its validity and manipulability remain debated (Morgeson et al. 2007). We combine survey measures with the firm's administrative records and train random-forest models on employees with at least 12 months of tenure, a horizon that managers view as sufficient for reliable performance assessment. Our primary outcome is a pre-registered binary productivity metric defined by the firm: whether employees qualify for a bonus within the first year, based on portfolio size and repayment performance.

Within the sample of employees with at least 12 months of tenure, we use repeated holdout validation, i.e., repeatedly training the model on one randomly selected subset of employees and evaluating it on the remaining employees, to assess predictive performance. A random forest model using only firm administrative data, such as age, education, marital status, region, and

[3]In addition, even accurate tools may face implementation constraints because managers, clients, and workers can be averse to algorithmic decision-making or override recommendations (Highhouse 2008, Dietvorst et al. 2015, Dargnies et al. 2026).

related personnel characteristics, correctly classifies 65.1% of employees, on average, as bonus-eligible or not. This exercise evaluates predictive accuracy on randomly held-out employees drawn from the same underlying sample. Adding the non-incentivized measures significantly improved the performance of the random forest algorithm by five percentage points. A similar improvement was observed with the use of incentivized measures. However, the inclusion of incentivized measures alongside non-incentivized ones did not further enhance the algorithm's performance, suggesting that the two sets of measures are substitutes. Because non-incentivized measures are easier to elicit in practice, the remaining analyses and the field experiment use the algorithm trained on firm data and the non-incentivized measures.

A first concern in applying the survey-augmented algorithm to hiring is that the training sample reflects prior HR choices and subsequent retention outcomes, which may distort model performance through selective labels (Kleinberg et al. 2018, Chalfin et al. 2016). As a second step, we partially address this problem and provide an out-of-sample test that accounts for on-the-job selection. This exercise relies on the sample of employees with less than one year of tenure at the time of the survey. Compared to the training sample, this sample is less selective, as it includes recent hires—some of whom may leave the firm within a year—yet it remains selective insofar as it is restricted to applicants who were hired. We find that those predicted to be productive by our algorithm are significantly more likely to receive a bonus, exhibit higher productivity, manage larger portfolios and issue more loans at the 12-month tenure mark. They are also significantly less likely to leave the firm within the first year of employment, consistent with positive selection in the workplace. These results suggest that, in this setting, selection into the long-tenure training sample does not generate large distortions in the algorithm’s categorization of workers.

Another concern regarding survey-augmented hiring algorithms is that strategic responses invalidate survey measures. In the third step, we conducted a randomized field experiment on hiring new employees with three goals. First, we evaluate the effectiveness of algorithmic hiring compared to the firm’s current hiring practices. Second, we further address the selective labels problem by predicting job applicant performance under alternative hiring regimes—one based on AI recommendations and the other on the traditional HR process. Third, the experiment allows us to assess whether potential manipulations of survey responses by applicants affect the algorithm's accuracy. For one year, all applicants to the firm completed a short version of the survey during their interview process in the local offices. This survey provided the inputs for the non-incentivized measures used by the algorithm. We randomly assigned applicants to the HR

and AI treatments. In the HR treatment, local and regional managers made the hiring decisions as usual, but we also recorded the algorithm's recommendation. In the AI treatment, a decision by local and regional managers was overridden if it conflicted with the algorithm's recommendation.

To address the first goal of the experiment—evaluating the effectiveness of algorithmic hiring relative to the firm's HR practices—we compared employee performance under the two experimental treatments at the 12-month mark. During the experiment, managers rejected far fewer candidates than anticipated, which reduces treatment variation and limits statistical power of the experiment.[4] Thus, we are comparing a sample selected by AI with an almost non-selected sample. Employees hired in the AI treatment were 9 percentage points more likely to receive a bonus and were 0.31 standard deviations more productive after one year. The corresponding Romano–Wolf adjusted p-values are 0.07 and 0.06, and we interpret these treatment effects as marginal after controlling for the family-wise error rate. Beyond performance indicators such as bonuses and productivity, a key outcome for the firm is the profitability of newly hired employees. Profitability shows a clearer pattern than productivity: average profitability at 12 months is significantly higher in the AI treatment, and despite fewer hires overall, the total profit generated by AI-hired employees over the year by far exceeds that of the HR treatment. Taken together, these findings provide supportive evidence that algorithmic hiring improved realized hire quality. The estimated effects on the firm's bonus and productivity margins are economically meaningful and conventionally significant, but only marginal after Romano–Wolf adjustment across the outcome family. The profitability results point in the same direction.

To address the second and third goals of the experiment, i.e., assessing how harmful selective labels and strategic answers are, we compare applicants whom the algorithm recommended for hiring with those it did not. Since managers in the HR treatment accepted nearly all candidates, we observe a substantial number of individuals whom the algorithm would have rejected but who were nevertheless hired. Although the unexpectedly low rejection rate reduces experimental variation in hiring decisions, it provides a key advantage for evaluating the algorithm: Assessing the predictive performance of an AI hiring tool requires observing job outcomes for representative samples of candidates the algorithm would accept and candidates it would reject.[5]

[4]The top management expressed surprise at this outcome and hypothesized that the low rejection rate might be attributed to an employee shortage throughout most of 2022.

[5] As in any hiring context, a subset of selected candidates ultimately either do not join the firm or do not remain employed long enough to generate reliable performance measures. Importantly, attrition of this kind does not vary systematically with candidates' algorithmic recommendation status.

Applicants recommended by the algorithm were significantly more likely to receive a bonus, had higher productivity, and held portfolios with fewer delayed loans at the 12-month mark relative to those not recommended. In contrast, we find no robust evidence of differences in retention at 12 months. Overall, these results indicate that the algorithm is robust to selection effects and to potential strategic manipulation of survey responses, although its predictive power for portfolio size, number of issued loans, and retention is weaker in the applicant sample than in the employee sample.

We investigate whether applicants give strategic responses by comparing the distribution of responses to the survey of existing employees with candidates. Using propensity score matching on non-strategic variables such as education, age, numeric literacy score, and others, we find evidence of strategic responses in four out of 12 potentially strategic variables: patience, agreeableness, locus of control, and neuroticism.[6] At the same time, the proportion of applicants recommended for hiring by the algorithm does not increase over time in the one-year period of the experiment. If (successful) gaming occurs, this proportion should rise. Thus, while some signs of manipulation attempts exist, participants failed to learn how to manipulate the algorithm within the period of one year that the experiment lasted. A likely explanation is that the mapping from survey responses to hiring recommendations is difficult to infer for applicants, since many measures enter the prediction in a non-monotonic way. As a result, applicants cannot generally improve their chances by simply reporting what they perceive as more desirable values on each survey measure.

Finally, we quantify the value of behavioral measures as inputs to the algorithm relative to an algorithm trained solely on the firm's administrative data. Although the firm could in principle rely only on its own records—which are automatically collected and not subject to strategic responses—our results show that incorporating a compact set of non-incentivized behavioral measures improves predictive performance. When we generate parallel hiring recommendations using a firm-data-only algorithm, the two classifiers agree on most applicants, yet the disagreement cases are informative: employees recommended only by the algorithm with behavioral measures perform substantially better than those recommended only by the firm-data algorithm. Regressions including both recommendation indicators show that only the behavioral-augmented algorithm significantly predicts bonus attainment, productivity, and portfolio quality,

[6]We consider all answers to questions that measure behavioral traits as potentially strategic but assume that verifiable personal data such as education cannot be manipulated, as well as cognitive ability questions or tests that require correct answers (e.g., the RME test).

while the firm-only algorithm has no significant predictive power. Profitability results mirror these findings: both algorithms identify more profitable hires, but the behavioral algorithm yields a larger and more precisely estimated effect, implying that average profit per hire is more than 13 percent higher when behavioral measures are included. These patterns indicate that survey-based behavioral measures provide information that improves prediction beyond what the firm's administrative data alone can deliver.

Overall, our paper provides a constructive framework for training firm-specific hiring algorithms with a menu of potentially useful measures based on research in behavioral economics and psychology. The measures that prove effective may vary significantly depending on the context and tasks of employees. This contrasts with the common practice of firms offering algorithmic hiring services where "one-size-fits-all" algorithms are employed to assess candidates. In our setting, survey-based behavioral measures add economically meaningful information beyond administrative records, while selective labels and strategic responses—though present—do not eliminate the effectiveness of the approach.

Related literature

We build on evidence that personality traits, preferences, and behavioral measures predict or affect important economic and life outcomes such as wages, health, and longevity (Heckman et al. 2019), portfolio choice and self-employment (Dohmen et al. 2011), career sorting and academic performance (Gottfredson 2002), earnings (Dohmen et al. 2009), school-track and university choices (Buser et al. 2014; Hakimov et al. 2023), and educational attainment (Alan et al. 2019; Hanushek et al. 2023); surveys and meta-analyses document their relevance for lifetime earnings and other life outcomes (Bowles et al. 2001; Kautz et al. 2014; Beck and Jackson 2022). This literature largely studies individual measures in isolation, as correlates or causes of outcomes. We instead ask whether a bundle of such measures can improve predictions of employee performance and remain useful when applicants know the measures may affect hiring.

We also relate to work on the relationship between psychological and economic measures (Dean and Ortoleva 2019; Jagelka 2024). We compare alternative combinations of economic, psychological, and cognitive inputs and show that non-incentivized measures from economics and psychology complement one another in predicting productivity. The field experiment further lets us compare robustness to strategic manipulation, yielding suggestive evidence that economic measures are harder to manipulate than Big Five traits.

Our paper contributes to research on tests, machine learning, and AI in personnel decisions. Hoffman and Stanton (2024) review recent work on technology and HR practices. Awuah et al. (2025) compare human screening, AI-augmented human screening, and fully automated ChatGPT-4.0 screening in teacher hiring and find that automated evaluation predicts subsequent performance better. Autor and Scarborough (2008) show that job testing improves selection without reducing minority hiring. Hoffman et al. (2018) show that managers who hire against test recommendations select applicants with lower retention. Chalfin et al. (2016) demonstrate the potential of machine learning in a data-rich public-sector setting. Relative to this work, we study a firm with limited administrative data and evaluate a firm-specific hiring algorithm in a randomized field experiment.

Several related papers study the labor-market and diversity effects of AI hiring. Avery et al. (2023) find that AI hiring increases women's applications in a stereotypically male profession and improves managers' evaluations of women when AI scores are available. Awad et al. (2023) show in online experiments that AI does not change applicant quality or gender diversity relative to human evaluators, while debiasing humans or algorithms improves gender diversity without reducing high-quality applicants. Zhang and Kuhn (2024) audit job recommendations and find gender differences in recommended jobs and wages. Jabarian and Henkel (2025) show that AI-conducted interviews with human scoring increase job offers, job starts and workers retention through more structured interviews with no decline in the productivity of hired workers. Li et al. (2026) find that exploratory algorithmic hiring selects better applicants and more good candidates from underrepresented groups. Work on acceptance of AI hiring is mixed. Lee (2018) and Kaibel et al. (2019) document reservations about algorithmic hiring, whereas Bigman et al. (2023) and Corgnet (2023) find less negative reactions to algorithmic than to human decisions.

Another strand studies strategic adaptation to profiling and testing. Bonatti and Cisternas (2020) show theoretically that welfare effects of consumer scoring depend on consumers' manipulation strategies; Bó et al. (2024) find that participants can alter survey responses when lottery-ticket prices may be personalized; and Hagenbach and Salas (2025) show that most participants fail to optimally conceal information from an algorithm. Psychology research reaches related conclusions for personality measures. Viswesvaran and Ones (1999) show that instructed faking raises Big Five scores. Birkeland et al. (2006) document differences between applicants and non-applicants, while acknowledging the potential role of selection. Roulin and Krings (2020) show that applicants tailor their personality profiles to align with perceived organizational fit. We contribute by studying strategic responses in an actual survey-based hiring process using both

psychological and economic measures, and by comparing applicants with employees while controlling for cohort and selection differences.

Our paper also relates to a literature showing that behavioral economics can inform the design of firm policies. Hossain and List (2012) show that framing bonuses as losses rather than gains increases productivity in a Chinese factory, while Gosnell et al. (2020) find that personalized feedback and goal setting improve fuel efficiency among airline captains. Kaur et al. (2015) show that commitment devices can reduce procrastination and raise workplace productivity. Other studies emphasize the social and organizational context in which firm policies operate: interventions targeting social norms and communication improve workplace climate, satisfaction, and engagement (Alan et al. 2023); management practices are more effective when they are aligned with employees' perceptions and intrinsic motivations (Blader et al. 2020); and incorporating worker evaluations into managerial decision-making raises productivity in auto manufacturing (Cai and Wang 2022). Relatedly, fairness concerns and reference points shape the implementation of payment reforms (Krueger and Friebel 2022), and employee referral programs are effective partly because workers value being involved in the hiring process (Friebel et al. 2023). We contribute to this literature by shifting attention from behavioral interventions within the workplace to behavioral measurement at the hiring stage. Our results show that standardized behavioral measures can help firms identify productive workers and improve AI-driven hiring decisions, especially when the firm's own administrative data are limited.

Finally, we speak to the literature comparing incentivized and non-incentivized measures. Dohmen et al. (2011) compare survey risk attitudes with incentivized responses. Vieider et al. (2015) replicate this relationship across 30 countries. Lönnqvist et al. (2015) compare a survey-based risk measure with an incentivized task developed by Holt et al. (2002) and find the survey question more stable. Falk et al. (2018) develop non-incentivized measures for economic preferences, and Hackethal et al. (2023) find that incentives do not significantly affect risk-elicitation results. Chapman et al. (2025) challenge the validation of qualitative self-assessments, arguing that they correlate with multiple incentivized measures and can be difficult to interpret. Our results show that non-incentivized survey measures can substitute for incentivized measures in prediction, although they may require more questions. We remain agnostic about what these qualitative measures capture structurally and use them as proxies for productivity.

2 Research design

We collaborate with a microfinance company in Kyrgyzstan. The study focuses on the loan officers of the company whose main job is to find clients, evaluate their creditworthiness, and monitor repayments of loans. The work of loan officers is complex, and the firm's management lacks a clear profile of what defines a successful employee. This can in part be explained by the dual role of specialists: they need to sell a high volume of loans while being selective in targeting only creditworthy clients. During the study, the specialists could issue loans ranging from approximately $100 to $3,000, with an average interest rate of 31%. The maximum loan and the individual interest rate were determined automatically based on the client's credit history within the firm. Loan officers did not have any influence over the interest rate, nor could they issue amounts exceeding the limit. However, they had the right to either reject the entire loan or approve only part of the requested amount. To determine creditworthiness, loan officers evaluate the economic conditions and the repayment potential of prospective clients, but this dimension of their productivity only becomes evident over time, as repayment challenges typically arise three to four months after loans are issued. The firm faces significant heterogeneity in the productivity of their employees and experiences high turnover among those recently hired, many of whom leave just as repayment issues start to emerge. While we know who leaves the company at what point in time, we do not observe whether a person was fired or quit.

A distinctive feature of our partner firm is the systematic tracking of each loan officer's performance through a measure that we call "productivity". Productivity is an internal performance measure proportional to the monthly profit generated by each officer. It is the main determinant of remuneration. Although the exact formula is proprietary, management confirmed that it is computed from the interest income earned for the officer's loan portfolio (positive component) and the value of delayed or delinquent loans (negative component). The calculation is performed at the head office and is fully transparent to each employee, who can monitor the monthly figures in the firm's internal system. Importantly, while the formula is uniform and objective, employees do not observe the bonuses of their peers.

Whenever an officer's productivity exceeds their fixed salary, the difference is paid as a bonus. The incentive intensity of this system is high: on average in 2023, the total amount of bonuses paid to loan officers was 2.7 times higher than total salaries, and top performers could receive bonuses up to the equivalent of 25 monthly base salaries of around $200. This strong pay–performance link generates substantial heterogeneity in earnings and career trajectories, with

productivity strongly increasing in tenure due to returning clients who obtain larger loans and are lower risk. Conceptually, the firm's compensation scheme reflects a standard principal–agent framework in which performance-based pay aligns employee incentives with the firm's objectives.

As pre-registered, we focus our analysis on a binary indicator of whether an employee received a bonus within the first year. This choice is motivated both by the firm's internal definition of success—qualifying for a bonus or not—and by empirical considerations of power, since predicting a binary threshold of high performance is more feasible than forecasting a continuous productivity measure in a modest sample.

The study consists of three stages

Stage 1. Collecting behavioral measures and training the algorithm

In this first stage, employees of the firm as of September 2021 answer survey questions to elicit their preferences, cognitive skills, and psychological traits. Some of the questions are incentivized while others are not. The employees are informed that the purpose of the survey is to collect data to improve the management of the company and that their individual responses will not be known by any of their peers or by the local and regional managers.

In September 2021, all current 1042 employees answered the survey. The average duration was one hour and six minutes. The survey was administered in Russian and Kyrgyz, the two main languages spoken in the firm, and participants could choose the language. The remuneration of the incentivized questions was paid out with the salaries. Only one of the incentivized measures was randomly drawn for each employee to determine the payment.

We measure the employees' risk and time preferences, trust and trustworthiness, altruism, the Big 5 personality traits, performance in the Cognitive Reflection Test (CRT), a numeric literacy test, the Wonderlic Test, and the Reading the Mind in the Eyes Test as well as self-confidence. We elicited in total five incentivized measures and 22 non-incentivized measures. The complete list of measures and corresponding questions is presented in Appendix A.

We anonymously match the survey responses to the personnel data of the firm that include measures of productivity of the employees. These measures are the portfolio, the portfolio at risk, the portfolio without delayed payments, the number of new loans issued and whether the employee qualified for a bonus in addition to their fixed salary. The latter is directly related to

an internal measure of individual productivity. The primary goal of the management when selecting new employees is to identify those who will qualify for a bonus, and the number of salespeople in a local manager's office who have obtained a bonus is a performance indicator. The formula for calculating productivity and the monthly bonus is transparent for all employees.

We train an algorithm to predict which employees perform best using those employees who have been working at the firm for at least one year as of September 2021. Of the 1042 employees, 674 were employed by the firm for 12 months or more. Based on management's assessment that one year is typically enough to qualify for the company bonus, we use these 674 employees to train the algorithm. We train the AI, a random-forest algorithm, to predict the pre-registered binary outcome of whether an employee is high achieving, defined as receiving a bonus within one year of employment. Algorithms predicting other variables (longer-term performance, turnover, portfolio at risk, size of portfolio) are run for the sake of exploration.

For the prediction of our main outcome variable, i.e., the payment of a bonus, one algorithm uses only firm data (such as age and education), another algorithm uses both firm data and the answers to the non-incentivized questions, and a third algorithm uses firm data, the answers to the non-incentivized questions, and to the incentivized questions. Since non-incentivized and incentivized measures turn out to improve the algorithm and are substitutes, we choose the algorithm which uses firm data and the easy-to-elicit non-incentivized measures in the following stages.

Stage 2. Predicting the performance of employees with short tenure

The algorithm trained on firm data and the non-incentivized measures is employed to predict the performance of the employees who have been with the firm for less than one year when they answered the questionnaire. We assess their performance one year after the survey and evaluate the algorithm' predictions. This allows us to measure the out-of-sample efficiency of the algorithm. The training sample is biased since it consists of people who worked at the firm for at least one year. Thus, the algorithm's ability to predict the performance of employees with shorter tenure will reveal its robustness to this selection bias.

Stage 3. Using the algorithm for hiring decisions

Finally, we study the usefulness of the algorithm for actual hiring decisions. The field experiment has three goals. First, it enables us to evaluate the firm's current HR practices against algorithmic hiring. Second, we can test the robustness of the algorithm to sample selection. Third, the

experiment allows us to investigate whether potential strategic responses from candidates undermine the efficiency of the algorithm. For the experiment, some employees were hired following the normal procedure of the firm (applicants are interviewed by the local office manager and by a senior manager, i.e., the manager of a region) while others were hired following the recommendation of the AI algorithm.[7] The cutoff of the algorithm for hiring an applicant or not was determined based on the selectivity of the current HR procedure.

All job applicants between March 2022 and February 2023 were informed that they must answer the questionnaire online as a part of the screening process after their application for the job. The survey questions, only consisting of non-incentivized measures, were administered with the help of Qualtrics, using the applicants' smartphones. The average duration of the survey was 19 minutes. Whenever new applicants arrived at the firm, they were interviewed and asked to answer the survey questions. They were then evaluated by the algorithm which generated a recommendation whether to hire them or not. The algorithm was run on a computer of the researchers.

Normally, the hiring is done by the local managers (managers of the office), with the approval of senior management (managers of the region). Local managers were informed that there is a new step in the application--the survey--and that they will not be informed about the answers of the applicants. Managers continued to make their decisions based on interviews, without access to the survey responses. Thus, our experiment does not study whether the manager or the algorithm makes better use of identical information, but rather who makes better hiring decisions based on different sets of information. For the randomly determined half of the sample in the AI treatment, the recommendation was sent to the main manager of the front office, i.e., an executive board member, who communicated the decision to the local office through the regional managers. Independent of the recommendation of the local manager, this decision was implemented by the senior manager in this treatment.[8] For the other half of the applicants, i.e., those in the HR treatment, senior management followed their usual decision process without receiving the recommendation of the algorithm. Note that due to this design, for every applicant we know the hiring recommendation by HR *and* by the algorithm.

[7] There are three levels of managers: the local office managers, the managers of a region who are heads of several offices, and finally a board member who is responsible for the front office, the COO.

[8] It is not unusual that local managers' decisions are overruled. The head office checks the formalities and regularly rejects applicants, for instance, because of missing documents.

Neither the applicants nor the local managers were aware of participating in a study. Within the survey applicants were informed that their answers might be used for automated scoring, but they were not informed about the experimental assignment or whether the algorithmic recommendation would be binding. Only the board knew about the experiment. Applicants were notified of the outcome of their application—whether they were offered a position or not. The data from the survey were only available to the researchers.

The study was approved by the IRB of LABEX, University of Lausanne, and the legal team of the firm. The study was pre-registered at the AEA RCT Registry (AEARCTR-0008219).

3 Algorithm Development

We use a random forest algorithm to classify employees as those predicted to receive a bonus and those predicted not to receive it. Our training sample only includes those 674 employees who were at the firm for at least 12 months when they answered the survey in September 2021. We use an information gain (entropy) criterion for node splitting.

Model performance is evaluated using repeated holdout validation (also known as Monte Carlo cross-validation; Kohavi, 1995; Arlot and Celisse, 2010). Specifically, we draw 250 random splits of the data into 550 training observations and 124 validation observations. For each split, we fit the model on the training set and record validation accuracy on the held-out validation set (the share of correctly classified employees out of 124). Hyperparameters are chosen to maximize average validation accuracy across the 250 splits. In the final specification, the number of candidate predictors considered at each split is set to 10, and the forest contains up to 500 trees.

Our primary algorithm of interest uses the firm data and all non-incentivized measures. These measures are straightforward to collect and could be utilized for scoring applicants, without the financial cost of incentives and the potential organizational complexity of paying out the rewards for the incentivized measures. An algorithm with fewer variables reduces the survey length.

To determine the final set of non-incentivized measures included in the primary specification, we begin with a model that includes all available predictors. We then apply backward feature elimination guided by the model's split-based variable-usage importance (Guyon and Elisseeff, 2003): in each step, we remove the least important variable and re-estimate the model using the same validation protocol. We stop removing variables at the point where further elimination reduces average validation accuracy.

Using the final set of non-incentivized behavioral measures, we examine the incremental predictive value of (i) adding behavioral and psychological traits to firm data, and (ii) including incentivized measures. This allows us to quantify the trade-off between predictive performance and the feasibility of collecting additional inputs.

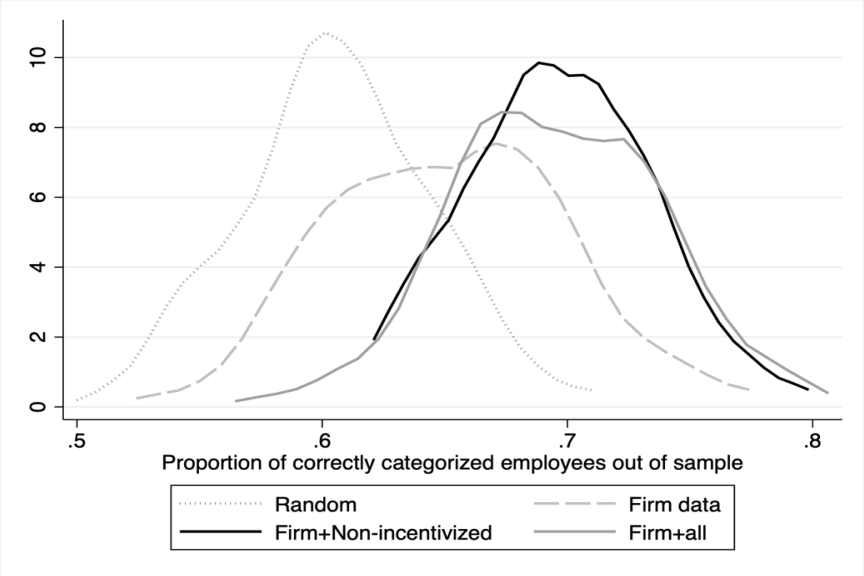

Figure 1: Validation accuracy of random-forest algorithms depending on set of variables used.
Notes: Accuracies are computed on the validation sample across 250 repeated random train/validation splits (550/124) and represent proportion of employees correctly categorized in the validation set of 124. The non-incentivized feature set corresponds to the final set selected via the backward elimination procedure described above. Hyperparameters (number of trees, candidate predictors per split, and any per-tree sample size setting) are tuned separately for each feature set using the same validation protocol.

Figure 1 presents the distribution of validation accuracy of random forest algorithms across the 250 repeated splits for models estimated on different sets of explanatory variables. Using firm data only, the random forest correctly classifies 65.1% of employees in the 124-person validation sample on average. This is significantly better than an algorithm based on ten randomly generated variables (a two-sided Fisher's exact test yields $p<0.001$). Adding the selected non-incentivized survey measures increases average validation accuracy to 69.7% ($p<0.001$). However, adding the incentivized survey measures to the algorithm that uses both the firm data and the non-incentivized survey measures does not further improve accuracy ($p=0.48$).[9] The firms' use of

[9] The incentivized survey measures also improve the algorithm's performance compared to the algorithm using only firm data ($p<0.001$). Comparing the performance of the algorithm using firm data and incentivized measures with the algorithm using firm data and non-incentivized measures, the algorithm with non-incentivized measures outperforms the one with incentivized measures but the difference is not significant ($p=0.11$). Note that the non-incentivized part of the survey contains many more measures than the incentivized part, such as cognitive skills, behavioral measures, and psychological traits.

incentivized measures may not be feasible for logistical and financial reasons, and the results show that this is not an obstacle to obtaining an algorithm with good predictive power.

We can also study whether the random forest algorithm allows us to correctly categorize a higher proportion of workers than a simple probit model. Figure 2 shows that when using firm data and the selected non-incentivized measures, the random forest achieves higher classification accuracy than the probit model ($p<0.001$). The magnitude of the effect is 1.5 percentage points, approximately one-third of the gain from adding the non-incentivized measures to firm data. A natural interpretation is that the random forest captures non-linearities and interaction effects among predictors that are not represented in the probit specification.[10]

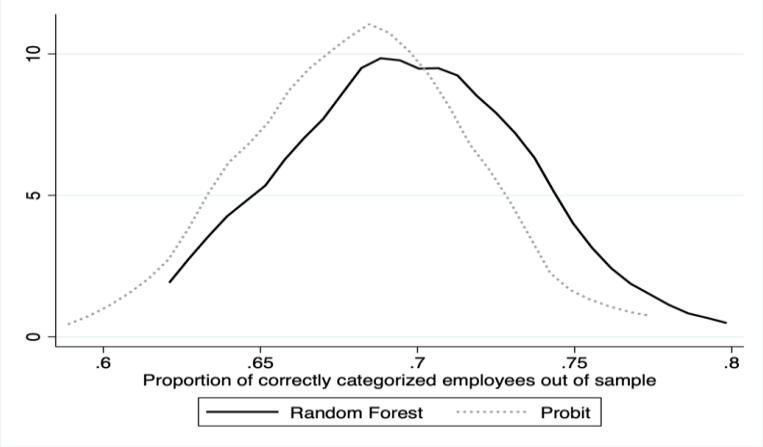

Figure 2: Validation accuracy of the random forest and probit regression. *Notes: Accuracies are computed on the validation sample across the repeated splitting protocol described above. The feature set corresponds to the final set of selected non-incentivized measures.*

Prior to the study, we agreed with the management that gender and ethnicity would not be used by the AI in the field experiment when deciding whom to hire. However, we can explore how adding these variables impacts the performance of the algorithms. Figure 3 shows that adding gender and ethnicity significantly increases accuracy when the model uses firm data only ($p<0.001$), whereas the improvement is not statistically significant when the model also includes the selected non-incentivized survey measures ($p=0.20$).[11]

[10] Figure B1 of Appendix B provides additional robustness evidence comparing algorithms based on firm data alone and firm data combined with non-incentivized measures. The density of model-predicted probabilities conditional on receiving a bonus confirms that non-incentivized survey measures improve predictive performance.

[11] There are two ways to interpret this. It is possible that gender and ethnicity are the main predictors of productivity, in which case the non-incentivized measures are predictive only to the extent that they themselves help to predict gender and ethnicity. Alternatively, it might be that the main predictors are captured by the non-incentivized measures, and gender and ethnicity correlate with these measures. To distinguish between these possibilities, we train two algorithms: one predicting the propensity to receive a bonus based on the gender and ethnicity of employees

Based on the performance of the algorithms and the impracticality of using incentivized measures, we choose the algorithm based on firm data and the selected non-incentivized measures. Table 1 presents the variables retained in the final algorithm and a split-frequency importance measure, averaged across the 250 repeated splits. The variables are ordered by importance, from highest to lowest. Importance is scaled relative to the most frequently used splitting variable, which is set to 100; all other variables are expressed as percentages relative to this benchmark.

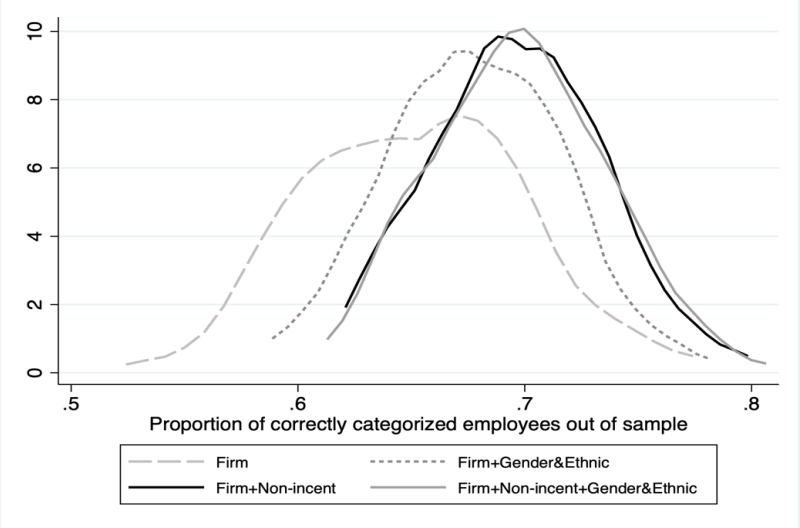

Figure 3: Validation accuracy of the random forest with and without gender and ethnicity as predictors.

The self-reported level of risk tolerance on a scale from 0 to 10 emerges as the most important variable, followed by self-reported patience on a scale from 0 to 10 and confidence in the number of correct answers in the Cognitive Reflection Test (CRT) and the numeric literacy test. Consequently, the three most significant variables are non-incentivized measures introduced by behavioral economists. The algorithm also uses some firm data, including regional effects represented by dummies for regions managed by different regional managers,[12] age, a dummy variable for holding a bachelor's degree, a dummy for specialization in economics or management, and a dummy for being unmarried. Notably, the final algorithm includes

only (average out-of-sample accuracy 59.5%), and the other based on the non-incentivized measures only (average out-of-sample accuracy 62.1%). The difference is significant ($p<0.01$). This suggests that the non-incentivized measures contain information beyond gender and ethnicity.

[12]The regions are an important predictor of productivity as they reflect different market conditions. Including them in the algorithm is beneficial for the firm. We do not assume that candidate quality varies by region, but using regional dummies allows the algorithm to adjust the hiring bar to regional differences, e.g., with respect to market structure and competition. In the capital region of Bishkek, where the bonus share is lowest, the algorithm requires a high score from observables to predict bonus eligibility. The relevance of each survey measure may also differ across regions due to differing client profiles. For instance, regions with strong bank competition might place more weight on employees' risk preferences, whereas this may be less critical in less competitive regions.

psychological measures, particularly extraversion, neuroticism, and agreeableness that are part of the Big Five personality traits. Additionally, locus of control and a subset of responses from the Reading the Mind through the Eyes test are included. We cannot determine the direction of a variable's effect on the probability of being classified as an employee who earns a bonus, since the relationship may not be monotonous.

Variable	**Importance**
Risk tolerance 0 to 10	100%
Patience 0 to 10	91.5%
Guessed number of correct answers	88.2%
Regional effects	80.7%
Age	78.2%
Trust: Assume best intentions	77.2%
N children	75.8%
Extraversion	74.1%
Numeric literacy	72.2%
Neuroticism	70.6%
Locus of control GSOEP	69.6%
Trust: Better be cautious with strangers	69.1%
Bachelor's degree	67.9%
Positive reciprocity 0 to 10	67.7%
Locus of control Rotter scale	67.6%
Agreeableness	65.6%
Specialization Econ or Management	65.2%
Altruism 0 to 10	64.7%
RME test	64.6%
Single (not married)	64.4%

Table 1: Importance of variables in the final random forest algorithm with firm data and non-incentivized measures. *Notes: Risk tolerance 0 to 10 is the response to "How willing or unwilling you are to take risks?"; Patience 0 to 10 is the response to "Would you describe yourself as a patient person?"; "Guessed number of correct answers" refers to the belief of the number of correct answers in the Cognitive Reflection and Numeric literacy tests; "Regional effects" are dummies for the different regions with their own regional managers; "Trust: Assume best intentions" refers to agreement to "As long as I am not convinced otherwise, I assume that people only have the best intentions."; "Extraversion" is measured based on five questions of the Big Five personality test; "Numeric literacy" is the number of correct answers to the numeric literacy test; "Neuroticism" is measured based on five questions of the Big Five test; "Locus of control GSOEP" based on seven questions used in GSOEP; "Trust: Better be cautious with strangers" captures agreement to "When dealing with strangers it is better to be cautious."; "Reciprocity" captures agreement to "How willing are you to return a favor if someone did you one?"; "Locus of control Rotter Scale" is the abbreviated 4-item Rotter Internal-External Locus of Control Scale; "Agreeableness" is measured based on five questions of the Big Five personality test; "Altruism 0 to 10" captures agreement to" How willing are you to give to good causes without expecting anything in return?"; "RME test" stands for performance in three questions of the Reading the Mind through the Eyes test which were selected based on the largest variation in the training sample.*

4 Predicting the performance of employees with short tenure

As demonstrated in the previous section, non-incentivized measures can be a useful input for predicting the productivity of employees. However, using the algorithm for applicants could suffer from two problems. The first concern is the selection of the training sample which only includes those employees who were hired and then stayed in the firm for at least 12 months. The algorithm could not be trained on data from applicants who were not hired or employees who left the company before completing one year of tenure. As a result, the algorithm's predictive accuracy may be compromised when applied to the broader pool of job applicants, limiting its effectiveness for hiring decisions. The second concern is that job candidates may answer some survey questions strategically to increase their chances of being hired, unlike the employees of the firm who answered the questions in September 2021. The two concerns are addressed in two steps. In this section, we limit selection effects by examining the future performance of employees with less than 12 months of tenure as of September 2021. The sample is less selective than the training sample, as it includes recent employees, some of whom may stay with the firm for less than one year, but it remains selective in the sense that it only includes applicants who were hired by the firm.

We test the effectiveness of the AI by predicting the performance of those employees who have been with the firm for less than one year in September 2021 when they answered the questionnaire. Overall, 368 employees were working at the firm for less than 12 months when they answered the questionnaire. The random forest algorithm based on firm data and non-incentivized measures predicts that 186 of them will get a bonus and 182 will not.[13]

Figure 4 shows outcomes by the algorithm's predictions. The left panel plots the proportion of employees who received a bonus at 12 months; the right panel displays the Z-standardized productivity measure among those still employed after 12 months. We pre-registered the measure of a bonus payment at 12 months of employment, but we also present results for productivity as a potentially more informative measure. We did not pre-register the productivity measure since the firm did not want to commit to sharing it with us at the design stage. After deliberations with them, we received a linear transformation of the productivity measure that preserves the salary information but allows us to analyze a cardinal measure of productivity conditional on

[13]Note that in the training sample of employees with at least 12 months of tenure, the algorithm predicts that 69% of employees receive a bonus, compared to 50.5% for employees with less than 12 months of tenure. This may reflect positive selection among longer-tenured workers.

employment. Employees predicted to receive a bonus are significantly more likely to do so than those predicted not to (two-sided Fisher's exact test, $p < 0.01$). Among employees still employed at 12 months, those predicted to receive a bonus also exhibit significantly higher productivity (two-sided t-test, $p < 0.01$).

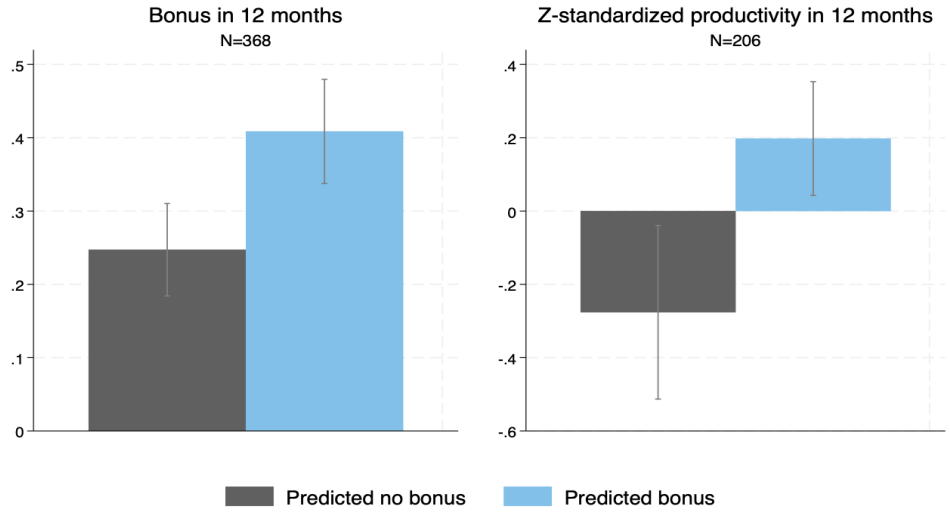

Figure 4: Bonus and productivity, conditional on employment, depending on the prediction of the algorithm.
Notes: Gray lines represent 95% confidence intervals. Sample of employees at the firm for < 12 months when responding to the survey.

Table 2 summarizes regression results for all five pre-registered outcomes plus productivity. The third-row reports Romano–Wolf p-values that adjust for multiple outcomes. On top of differences in the propensity to receive the bonus and productivity, those predicted to receive a bonus have larger portfolios and issue more loans within a year; they are also significantly less likely to leave within the first year. The difference in the size of the portfolio with delays is negative but not statistically significant.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
Algorithm predicts bonus	1.2e+06***	28.16***	0.159***	-0.165***	-4.0e+04	0.474***
	(3.7e+05)	(7.83)	(0.05)	(0.05)	(2.6e+04)	(0.14)
Romano-Wolf p-value	0.01	0.01	0.01	0.01	0.13	0.01
Observations	368	368	368	368	206	206
Sample	All	All	All	All	Employed	Employed

Table 2: Performance and turnover of employees with tenure < 12 months in 1 year. Coefficients of OLS regression for portfolio size, issued loans and productivity, and marginal effects of probit regressions of the bonus and left dummies on the AI predicts bonus dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To sum up, the algorithm has strong predictive power when forecasting the future performance of employees with short tenure. This suggests that any bias arising from the training sample, which consists of employees with longer tenure, is minimal in our context. Nevertheless, employees with shorter tenure still constitute a selected sample, as they were hired by HR, unlike the job applicants to whom the algorithm is intended to be applied. In the following section, we present the field experiment conducted to test the robustness of the algorithm when applied to job applicants and their potentially strategic responses.

5 Design of the field experiment

The field experiment serves to evaluate the effectiveness of algorithmic hiring compared to current HR practices. Starting in March 2022, all job applicants completed a pre-interview survey hosted on the University of Lausanne's Qualtrics server. The survey only included the questions used by the algorithm based on firm data and data from the non-incentivized survey. On average, the survey took 19.5 minutes to complete. Access to the survey answers was restricted to the researchers, and applicants were informed that their answers might be used for automated scoring. Every two to three days, we communicated the algorithm's hiring recommendations to the head office.

Two treatments were implemented:

(i) In the AI treatment, the recommendation of the algorithm whether to hire the applicant or not was implemented regardless of the recommendation of the management.

(ii) In the HR treatment, the recommendation of the management was implemented regardless of the recommendation of the AI.

Neither local nor regional managers were aware of the study. Top managers at the head office in Bishkek were our only contact points. They received the AI recommendation in the AI treatment, and communicated it to the local managers. In general, it is not uncommon that local managers' decisions are overruled by the head office, for example because there are concerns about the legal status of an applicant. As a result, the involvement of the head office and the transmission of recommendations did not constitute a departure from standard practice, making it unlikely that local and regional managers suspected they were part of an experiment.

To set an appropriate cutoff for the AI's hiring recommendations, we consulted with the management of the firm to determine the target rejection rate of applicants. While the firm does

not keep track of rejected applicants, they told us that they reject around 30% of the applications. We agreed that the algorithm would be calibrated to target a 30% rejection rate. We used the data from existing employees (the training sample and the sample of employees with short tenure) to determine the threshold for rejections.[14] For hiring, for each run of the model, we used a random subsample of 550 out of 674 current employees with at least one year of tenure. Each applicant was scored 250 times under different algorithms trained on a random subset of the training data. Thus, each applicant received a score between 0 and 250, depending on how many of the 250 algorithms predicted the candidate would receive a bonus. The threshold for hiring was based on the number of times the algorithm predicts the employee will receive a bonus. We calibrated this threshold to reject 30% of candidates based on the sample of employees with less than 12 months of tenure.

Between March 2022 and February 2023, every applicant was evaluated by both the AI and HR. Overall, 1183 applicants completed the survey between March 2022 and February 2023, 590 in the AI treatment and 593 in the HR treatment. The algorithm recommended hiring 63.7% and 62.6% of applicants in treatments AI and HR, respectively (p=0.72).[15] The managers recommended hiring 96.6% and 97.8% in treatments AI and HR, respectively (p=0.22). Thus, the treatments were balanced with respect to predicted performance, but the rejection rate by the managers was well below the expected 30%. This came as a surprise to the top management, and they hypothesized that the low rejection rate might be due to a shortage of employees in 2022. For our experiment, the fact that the local and regional managers recommended to reject only very few applicants means that we have less variation across treatments than expected. Only a small number of applicants hired in the AI treatment would have been rejected by the managers. Thus, almost all treatment variation is due to 23% of the hired sample being applicants recruited in the HR treatment who would have been rejected by the AI.

[14] To minimize sample bias relative to new applicants, we also include employees with short tenure in the threshold calculation.

[15] Note that the resulting rejection rate is higher than the target rejection rate of 30% due to differences between the pool of existing employees used to calibrate the rejection threshold and the pool of applicants.

In total, 957 applicants received an offer in the experiment, which corresponds to 81% of applicants. Of those applicants, 44% (421 applicants) ended up not joining the firm with 44.4% and 43.6% (p=0.84) of applicants in the AI and HR treatments, respectively.[16] Finally, 536 applicants ended up being hired by the firm. Of these 536 applicants, 327 were hired in the HR treatment of which 62% were recommended by the algorithm. In addition, 209 applicants were hired in the AI treatment, 96.7% of them recommended by the managers.

6 Results of the field experiment: Algorithmic versus HR hiring

6.1 Treatment effects

As pre-registered, we examine the full set of outcomes: (i) portfolio size, (ii) number of loans issued, (iii) the portfolio with delayed repayments, (iv) probability of leaving the firm within one year, and (v) probability of receiving a bonus.[17] In addition, we obtained access to a continuous measure of *productivity*,[18] which determines the bonus amount, a figure the firm was not prepared to share with us at the design stage. Note that portfolio with delayed repayments and productivity are measured conditional on continued employment after one year, while all other outcomes are observed regardless of employment status.[19]

Figure 5 illustrates the treatment effects on bonuses and Z-standardized productivity. The left panel shows that the proportion of employees who received a bonus is higher among those hired in the AI treatment than among those hired by managers (two-sided Fisher exact test, p = 0.02 for the probability of receiving a bonus at 12 months). The right panel presents the average Z-standardized productivity conditional on employment after 12 months. Employees hired under the AI treatment exhibit significantly higher productivity than those in the HR treatment (two-sided t-test, $p < 0.01$).

[16] This is most likely driven by a salary offer that is lower than expected. Candidates only learn about the exact conditions of employment during the interview stage.

[17] We pre-registered those measures at 6 and 12 months of tenure. The 6 months results are in Table B1 of Appendix B. In our context, productivity differences emerge with a delay due to close monitoring in the first months of employment and to lagged realized risk of loans.

[18] We have access to the Z-standardized measure of productivity which preserves secrecy of salary information.

[19] The portfolio is assumed to be 0 for employees who left the company.

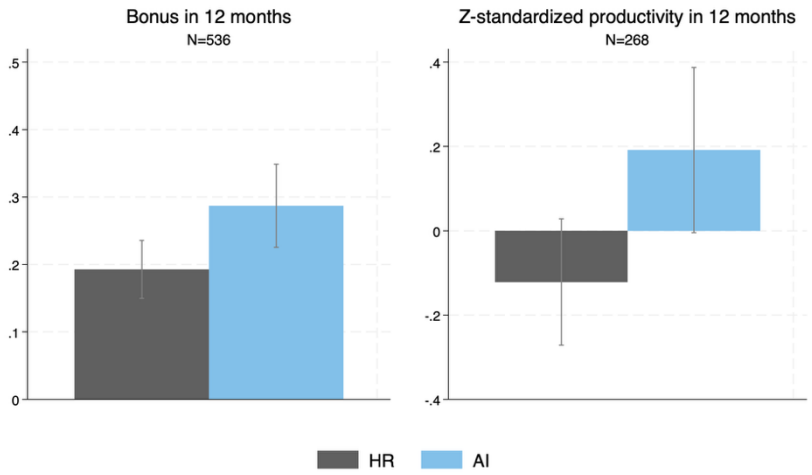

Figure 5: Treatment effect on bonus and productivity.
Notes: Gray lines represent 95% confidence intervals. Sample of all applicants hired in left panel; applicants still employed after 12 months in right panel.

Table 3 reports treatment effects for all outcomes based on regression analyses.[20] The third row presents p-values adjusted for multiple comparisons using the Romano–Wolf correction. Employees in the AI treatment are significantly more likely to reach the bonus threshold and are substantially more productive conditional on employment after one year. However, when accounting for multiple-hypothesis testing, the significance of both outcomes becomes marginal (p = 0.07 and 0.06, respectively). All other outcomes are statistically indistinguishable between the AI and HR treatments.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
AI treatment	3.1e+05 (2.9e+05)	9.90 (9.65)	0.09** (0.04)	0.004 (0.044)	-9.8e+03 (1.6e+04)	0.313*** (0.124)
Romano-Wolf p-value	0.63	0.63	0.07	0.94	0.80	0.06
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

Table 3: Performance of employees. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays, and productivity, and marginal effects of probit regressions of the bonus and left-the-firm dummies on the AI treatment dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The mixed results regarding differences between employees hired by managers and by the AI can be due to the lack of treatment variation caused by the low rejection rate by the managers.

[20] In the preregistration, we announced analyses based on two intervals: 6 months and 1 year. Ex post, the one-year horizon turned out to be the more relevant interval for detecting meaningful differences, as repayment problems and most performance-related issues typically materialize later in the credit cycle. Nevertheless, for completeness and to remain consistent with the preregistered plan, we report the 6-month results in the Appendix (Table B1).

Our findings suggest that algorithmic hiring yields a slight improvement in efficiency over current HR practices in identifying productive employees.

6.2 Profitability of AI versus HR hiring?

A natural question is whether AI hiring was more profitable for the firm than traditional HR hiring. Profitability depends on several factors—hiring and training costs, managerial time, survey administration, and implementation costs of the algorithm—but these are not directly observable and difficult for management to estimate. Most importantly, profitability depends on the actual performance of employees.

We asked the firm to compute a simplified measure of profitability for every employee over the first 12 months, defined as total interest income from the employee's portfolio minus the value of delayed repayments and total salary payments, including social security costs paid by the firm. Because the underlying data include confidential information, we received a linear transformation of this measure that was not standardized to keep positive and negative values directly interpretable: positive values indicate profits, negative values indicate losses.

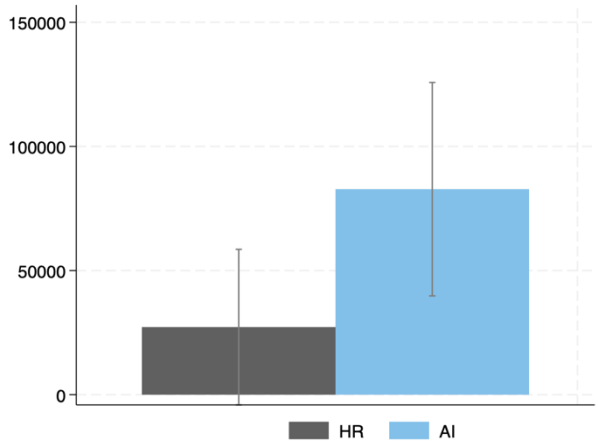

Figure 6: Treatment effect on profit per loan officer.
Notes: Gray lines represent 95% confidence intervals. Applicant sample with N=536. Profits are reported in transformed units obtained via a linear transformation of the underlying values to preserve confidentiality.

Figure 6 displays treatment effects on profit per loan officer. Average profit per employee is significantly higher in the AI treatment than in the HR treatment (two-sided t-test, $p = 0.04$). However, because more employees were hired in the HR treatment, total profit comparisons are also informative. The cumulative profit of all loan officers in the AI treatment amounts to 16.57 million units, compared to 8.01 million in the HR treatment. Hence, despite leading to fewer hires, the AI treatment generated substantially higher overall profits.

These findings suggest that algorithmic hiring not only improves average employee productivity but also enhances the firm's aggregate profitability, even over a relatively short time horizon. The low rejection rate by HR in the period we are considering may contribute to this result. In practice, one strength of AI-led hiring may be to avoid such failures to reject low-productivity applicants.

6.3 Potential bias under algorithmic hiring

A common concern with algorithmic decision-making is the possibility of implicit discrimination. Even when sensitive demographic variables are not used by the algorithm, correlations between behavioral measures and personal characteristics could lead to biased selection patterns. To assess whether the algorithm altered the demographic composition of employees hired, we compare potentially sensitive observable characteristics between the AI and HR treatments, namely age and gender.

The comparison shows no evidence that algorithmic hiring produced a gender or age bias. The proportion of female employees is slightly higher in the AI treatment (72%) than in the HR treatment (66%), but the difference is not statistically significant. Similar results apply to the age of the employees. The absence of gender bias is particularly notable given that behavioral measures—such as risk preferences and patience—can correlate with gender in other contexts. In this setting, their use did not translate into systematic gender differences among hired employees, indicating that the algorithm's predictions focused on productivity-related traits rather than demographic proxies.

Variable	Control (HR)			Treatment (AI)			Difference
	n	mean	sd	n	mean	sd	mean
Female	327	0.66	0.47	209	0.72	0.45	0.05
Age	327	27.74	7.27	209	28.23	6.35	0.49

Table 4. Balance in observable characteristics between HR (control) and AI (treatment) hiring. The standard errors are clustered at the office level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

6.4 Do local managers and the algorithm value the same skills?

The previous analysis compared the efficiency of algorithmic and HR hiring in selecting high-productivity candidates using different indicators of productivity. However, productivity alone does not define a good employee. It is crucial to consider whether the algorithm prioritizes high performance of employees at the expense of collegiality and teamwork. Relatedly, the AI may

hire applicants that local managers find unsuitable for reasons beyond productivity. Notably, the bonus system is purely based on individual performance, requiring no cooperation with colleagues. Yet, office atmosphere and team spirit might influence turnover. It is therefore important to assess how the employees selected by the algorithm compare to others in key non-performance skills.

To investigate this question, we administered a survey among local managers in January 2024. The aim of the survey was to obtain an informal evaluation of employees. We analyze whether the scores for collegiality etc., obtained with the survey, are correlated with the propensity of the employees to receive a bonus and with the algorithmic recommendation. The detailed explanations of the analyses and the table with the results of the OLS regressions are in the appendix C (Table C2). We find that the scores of employees provided by local managers are consistently higher for employees who receive a bonus than for those who do not. The scores are, however, not significantly associated with the recommendation of the algorithm. We conclude that although the algorithm is designed to predict a bonus payment, it does not select applicants with significantly weaker teamwork and social skills.

7 Validating the hiring algorithm and the role of behavioral measures

In this section, we ask how well the algorithm predicts candidate productivity. Two concerns motivate the analysis: (i) selective labels in the training data and (ii) strategic survey responses by applicants. We evaluate predictive power using outcomes from both the AI and HR samples. The AI arm delivers a non-selected set of hires among those recommended by the algorithm. Crucially, because HR hired (almost) all applicants, we observe outcomes for samples that are effectively representative of both candidates the algorithm would have recommended and candidates it would not have recommended. This provides post-hire outcomes for both recommended and non-recommended candidates, allowing us to assess predictive performance on an effectively unbiased set of new hires from the HR arm and to gauge the implications of selective training labels and strategic responses.[21] Finally, we isolate the contribution of behavioral measures by comparing an algorithm built solely on firm data with one that additionally incorporates behavioral variables.

[21] The remaining selection concern stems from self-selection into job offers, as 44% of candidates declined the offer. Reassuringly, candidates who declined do not differ from those who accepted in their probability of being recommended by the algorithm ($p = 0.83$). Under the assumption that selection on unobservables is not correlated with algorithmic recommendations, this pattern suggests that self-selection into employment is unlikely to bias our estimates of the algorithm's predictive performance in the subsequent analyses.

7.1 Algorithm predicting the performance of applicants

Since the algorithm was trained to predict the probability of receiving a bonus, we first test whether employees recommended for hiring by the algorithm are more likely to receive a bonus and exhibit higher productivity than those whom the algorithm did not recommend. Among the 536 applicants hired across treatments, 77 percent were recommended for hiring by the algorithm. Figure 7 shows that employees recommended by the algorithm are significantly more likely to receive a bonus at 12 months and display higher productivity conditional on employment (two-sided Fisher exact test for the probability of receiving a bonus at 12 months and two-sided t-test for productivity both result in $p < 0.01$).

We next examine all five pre-registered outcomes plus productivity. Table 5 reports the effects of the algorithmic recommendation across the full set of measures. The third row presents Romano–Wolf adjusted p-values to account for multiple hypotheses testing. Employees recommended for hiring by the algorithm perform better than those not recommended with respect to bonus payment, share of the portfolio with delayed repayments, and productivity (significance at the 5-percent level after multiple-hypothesis testing) while the difference for the remaining three measures is at best significant at the 10 percent level with corrections. These findings are not surprising given that the algorithm was trained to predict bonus payments and bonus payments are directly related to productivity. For reasons of power, we cannot perform the analysis on the HR and AI samples separately. But analogous regressions as in Table 5 that control for the treatment show similar results: Candidates recommended by the algorithm perform significantly better with respect to bonus, portfolio with delays, and productivity (Table B2 in the Appendix).

Overall, the random forest algorithm trained on firm data and non-incentivized behavioral measures predicts applicants' future performance well. Employees recommended by the algorithm are not only more productive with a higher chance of receiving a bonus but also manage portfolios with fewer delayed repayments. These results closely mirror the findings from the analysis of existing employees with shorter tenure, except that the algorithm does not significantly predict the probability of leaving the firm within one year. This could be due to some productive employees leaving the firm because they find a better job elsewhere. Unfortunately, we cannot investigate this further because we do not observe the reason for separation—whether an employee quit or was terminated—only that the employee left the firm.

We conclude that the algorithm's overall predictive accuracy is robust to both selection in the training data and potential strategic survey responses.[22]

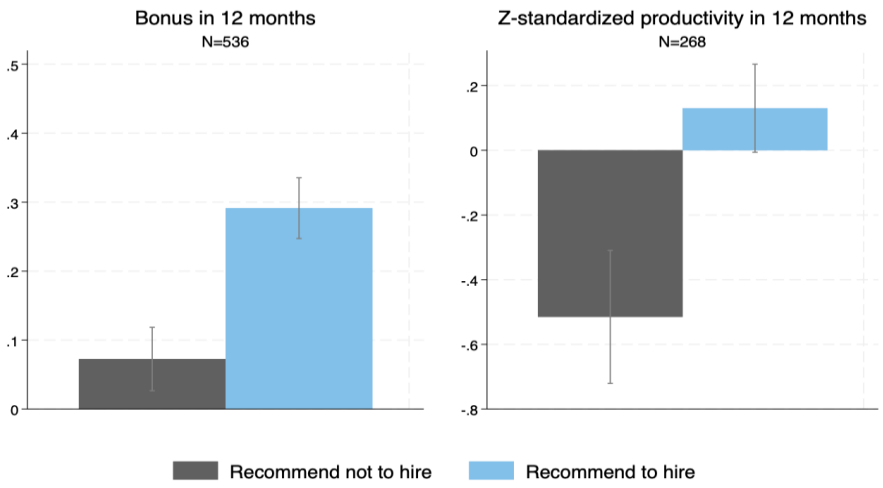

Figure 7: Proportion of employees who obtained a bonus and productivity conditional on employment, depending on the algorithm's hiring recommendation.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
Algorithm recommends hire	5.6e+05* (3.3e+05)	22.344** (11.134)	0.268*** (0.051)	-0.084* (0.051)	-7.8e+04*** (1.9e+04)	0.645*** (0.147)
Romano-Wolf p-value	0.13	0.06	0.003	0.13	0.003	0.003
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

Table 5: Performance of employees at 12 months. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays, and productivity, and marginal effects of probit regressions of the bonus and left-the-firm dummies on the algorithm recommendation dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7.2 Opening the black box of the algorithm's recommendation

We can identify which characteristics of job applicants increase the likelihood of being selected by the random forest algorithm, i.e., which behavioral measures influence the algorithm's prediction of a higher propensity to receive a bonus, as well as the direction of the effects. This is of stand-alone interest, but it will also allow us to check whether applicants hold accurate beliefs about how the algorithm works, enabling them to strategically adjust their responses.

[22] Note that the results become even stronger if we use continuous instead of binary predictions by the algorithm. The analysis is presented in Appendix C and summarized in Table C.1.

Figure 8 presents the coefficient plot for the hiring recommendation, where each input of the algorithm is used as an outcome variable and regressed on a dummy of whether the algorithm recommended hiring. All inputs were z-standardized to simplify comparisons; thus, the coefficients' magnitude can be interpreted in standard deviations. The inputs are presented in order of their importance for the random forest. First, those recommended for hiring by the algorithm are significantly less risk-loving than those not recommended, i.e., they report a 0.7 standard deviation lower risk tolerance on a scale from 0 to 10. Also, they report 0.5 standard deviations more patience on a scale from 0 to 10 and 0.6 standard deviations more confidence in the number of correct answers for the CRT and numeric literacy tests. We also observe that those recommended for hiring are 0.25 standard deviations more numerically literate, score lower on the neuroticism scale, and are more likely to have a bachelor's degree compared to those not recommended. All other inputs do not vary significantly between those who are recommended for hiring and those who are not, despite their importance for the algorithm. This indicates that these variables do not contribute to hiring recommendations through simple monotonic relationships For instance, age is the fifth most important variable for the algorithm, but the average age does not differ between those recommended to be hired and those who are not. This feature also helps explain why strategic manipulation is difficult. An applicant who attempts to improve their hiring prospects by reporting systematically "more desirable" values on survey measures is unlikely to succeed unless they understand the full decision rule of the algorithm.

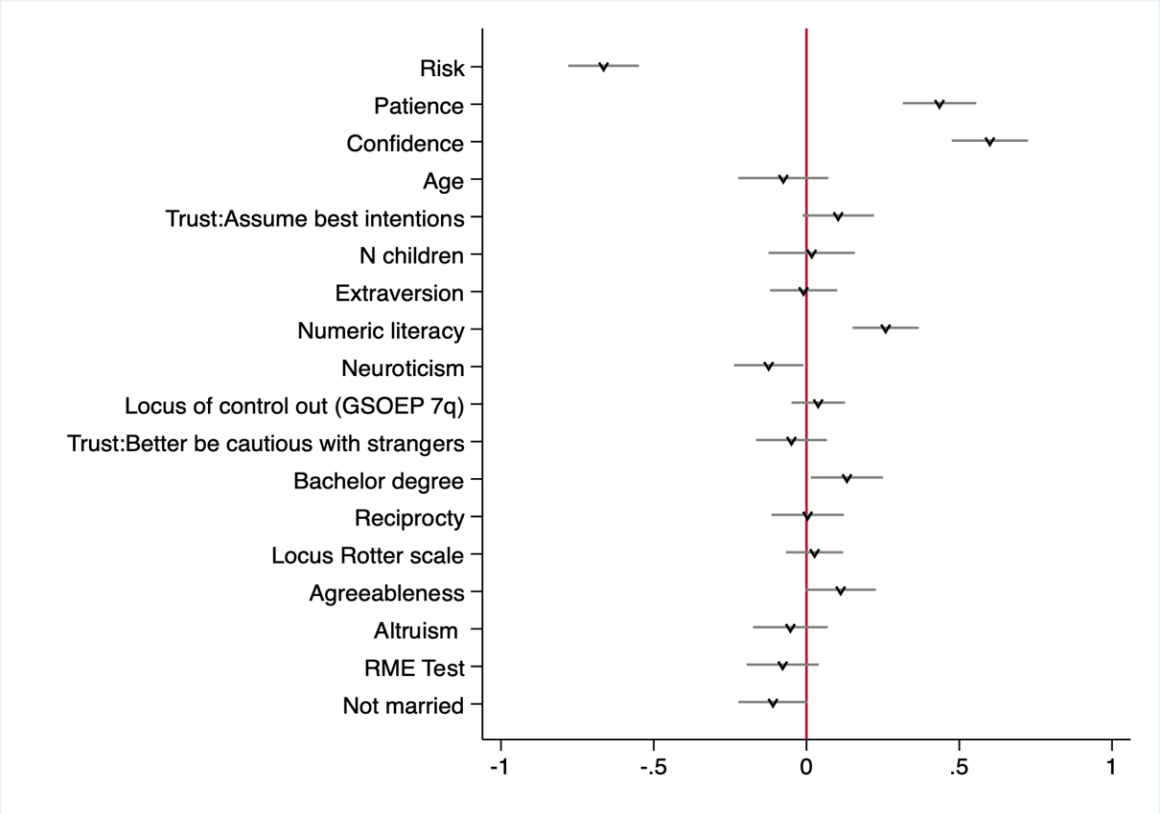

Figure 8. Coefficient of predicting recommendation to hire in OLS with the measure as outcome.
Notes: The red vertical line references zero. See notes of Table 1 for description of each measure.

7.3 Are responses strategic?

One of the main concerns regarding behavioral, non-incentivized measures as inputs for hiring is the potential for strategic responses from applicants who want to increase the probability of being hired. Candidates could manipulate their answers to the questionnaire to influence the algorithm's evaluation in their favor, leading to wrong hires and thus lowering the algorithm's performance. The results of Subsection 7.1 show that misrepresentations are limited, since the algorithm still distinguishes well between the applicants. However, our dataset allows us to identify which questions were more prone to strategic answers than others and in which direction applicants manipulated the responses.

We study the differences between inputs to the algorithm in the training sample, consisting of employees who participated in the survey in September 2021, and the sample of applicants from the field experiment. For the pooled dataset, we run OLS regressions of each input variable of the algorithm on a dummy variable indicating whether the individual is a job applicant (in which case the dummy is equal to 1) or an employee (dummy equal to 0). Figure 9 presents the coefficients of this "applicant" dummy in each regression: in Panel A without any controls and

in Panel B with controls. A significantly positive (negative) coefficient means that the input variable is higher (lower) for applicants than for employees. All inputs were z-standardized to simplify comparisons; thus, the coefficients' magnitude can be interpreted in standard deviations.

In Panel A, ten out of 18 inputs significantly differ between the employee and applicant samples. The differences in coefficients could be due to differences between the cohorts, e.g., due to selection or to strategic responses. For instance, the lower neuroticism of applicants compared to employees could be due to an actual difference in neuroticism between the two samples (selection) and/or to a successful attempt of applicants to appear less neurotic than they actually are (strategic responses). There are clear differences for some non-strategic variables, e.g., the applicants are significantly less likely to have a bachelor's degree than the employees and are more likely to be married.

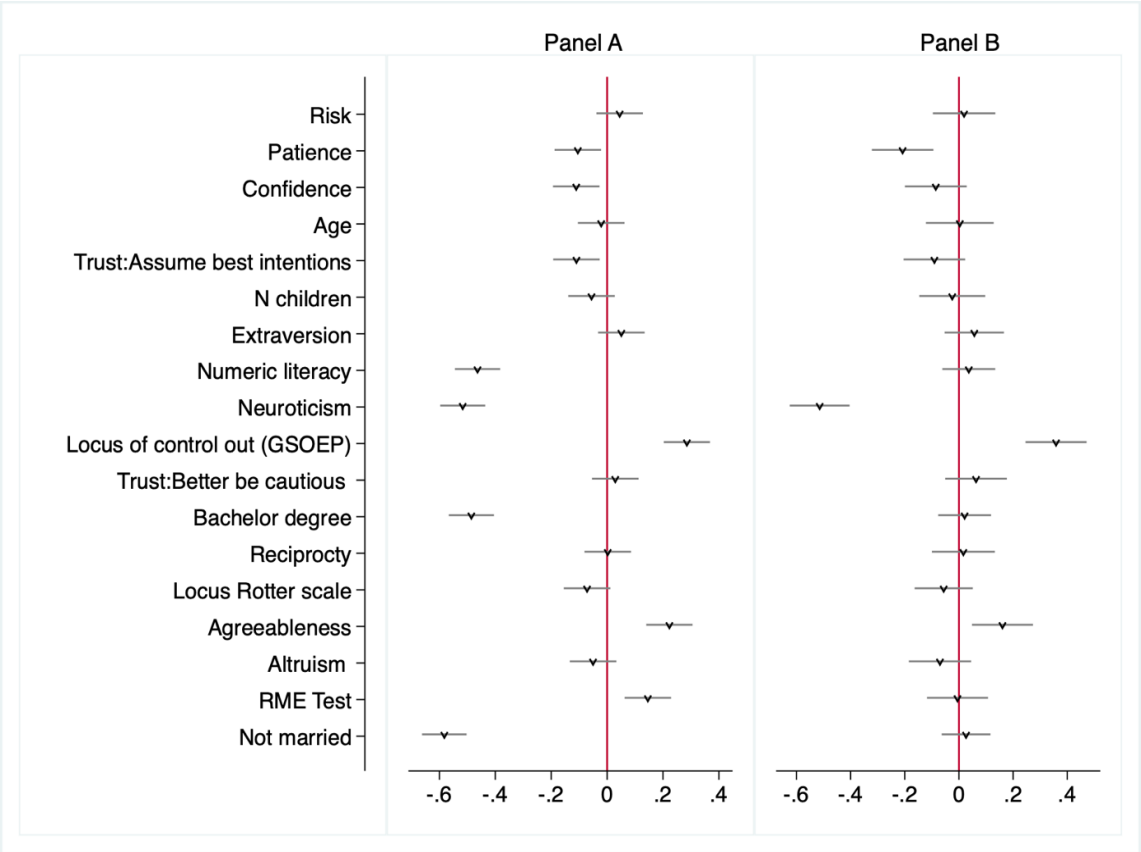

Figure 9. Coefficient plots of the OLS regression of the input variables of the algorithm on the dummy for the applicant sample relative to the employee sample.

Notes: Panel A presents coefficients for the dummy without any controls. Panel B presents coefficients for the dummy controlling for propensity scores based on age, gender, number of children, a dummy for bachelor's degree, a dummy for being single, the score in the numeric literacy test and CRT test, and the score in the RME test. The vertical line references zero. A negative coefficient means that the variable takes on a lower value for applicants than for employees. See notes of Table 1 for description of each measure.

While we cannot perfectly distinguish whether the differences are driven by differences in the composition of the samples or by strategic responses, we can get a better sense of it with the following exercise. In a first step, we calculate propensity scores of being in the applicant sample rather than the employee sample, based on a logit regression of a dummy for the applicant sample on age, gender, number of children, a dummy for whether the individual holds a bachelor's degree, a dummy for being single, the score in numeric literacy, the CRT[23], and the RME test. We selected these variables because the answers to them are likely to be non-strategic, assuming that applicants answer the questions of the CRT, numeric literacy, and RME test to the best of their ability. Then, we re-ran the OLS regressions from Panel A, comparing the values of inputs to the algorithm in the applicant and employee samples, adding controls for the propensity scores. Panel B of Figure 9 presents the results. As expected, fewer inputs differ between the employee and applicant samples than what we observed in Panel A without the controls. None of the non-strategic variables significantly differ, as expected when controlling for propensity scores. However, four inputs to the algorithm remain significantly different between the two samples. First, applicants report lower levels of patience compared to employees. This result is surprising as patience is valued by the algorithm (see Figure 8) and applicants should - if anything – portray themselves as more patient than they actually are if they want to increase their chances of being hired. One interpretation is that applicants perceive patience as an indicator of a lack of ambition. These patterns suggest that applicants may mistakenly believe that the algorithm rewards impatience rather than patience. The other three inputs are psychological measures. Applicants report lower levels of neuroticism (for instance, under-reporting the tendency to worry or get nervous), higher locus of control (the belief to be in control of events in life), and higher agreeableness than employees. These differences are consistent with applicants attempting to leave a good impression to improve their hiring chances. A lower score on neuroticism is expected to increase the likelihood of receiving a hiring recommendation, as shown in the previous subsection. For locus of control and agreeableness, however, the overall effect is zero, since it affects the likelihood of being recommended for hiring in a non-monotonic way.

Our finding of strategic responses to some of the psychological measures is not surprising, despite their prevalence in hiring practices. Studies in psychology have raised concerns about the manipulability of the Big Five traits (see, for instance, Roulin and Krings, 2020). Our results

[23]Note that the CRT variable is not included in the graph, since it did not make it into the final algorithm. However, we still use it in the propensity score matching to improve the model fit. Additionally, the measure of confidence refers to the participants' belief in how many correct answers they provided on both the numeric literacy test and the CRT combined.

complement these findings and indicate that economic measures are harder to manipulate, since their connection to productivity or employer preferences seems less predictable for candidates.

One might argue that manipulations are initially challenging, but candidates can learn to game the system. If this were the case, the proportion of candidates rejected by the algorithm should decrease over time. However, we see no indication of this during our experiment that lasted for one year. There is no upward trend in the proportion of candidates recommended for hiring by the algorithm each month, see Figure B2 in Appendix B. Additionally, none of the monthly fixed effects are significant (the lowest p-value being p=0.22 for June 2022 relative to March 2022). Thus, at least for the duration of our experiment, we do not observe any significant pattern of candidates learning to game the algorithm. We posit that the hiring algorithm is difficult to reverse engineer from applicants' perspective. While some characteristics exhibit clear directional relationships with hiring recommendations, others enter the prediction in non-monotonic ways (Figure 8). This suggests that the complexity of the prediction rule may itself provide some protection against strategic manipulation.

7.4 What is the role of behavioral measures?

How much of the algorithm's success is driven by behavioral measures rather than firm administrative data? In this subsection, we examine the incremental predictive power of behavioral measures for the future performance of newly hired loan officers. The firm could in principle rely only on its own data—although limited, these records contain predictive information on employee performance and have the advantage of being automatically available and not subject to strategic responses. The key question is therefore whether incorporating behavioral survey measures provides a measurable improvement in prediction accuracy and practical value for the firm.

To address this question, we generate an alternative set of hiring recommendations for the 536 newly hired employees using an algorithm trained solely on firm data, applying the same cutoff threshold for recommending a hire as in our main model that includes behavioral measures. Out of 536 loan officers, the two algorithms agree in 435 cases (378 hires and 57 rejections). Additionally, 34 employees are recommended for hiring by the algorithm with behavioral measures but not by the firm-data-based algorithm, while 67 are recommended only by the firm-data-based version.

We first regress all performance outcomes on both recommendations simultaneously. Table 6 reports the results of this measure of incremental predictive power of each algorithm. The algorithm that includes behavioral measures has more predictive power, significant for bonus payment, productivity, and portfolio quality. In contrast, the firm-data-only recommendation is not significantly associated with any outcome.

A complementary way to illustrate the importance of behavioral measures is to compare employee performance across the four possible agreement categories of the two algorithms. Figure 10 presents the shares of employees receiving a bonus within one year. Those recommended by both algorithms have the highest likelihood of obtaining a bonus, followed by employees recommended only by the algorithm with behavioral measures. In contrast, employees recommended by the firm-data-based algorithm but rejected by the behavioral algorithm have the lowest probability of receiving a bonus. This suggests once more that attempts by job applicants to answer some questions strategically do not critically harm the efficiency of the algorithm using behavioral measures.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
Behavioral algorithm recommends hire	3.6e+05	13.975	0.262***	-0.057	-8.0e+04***	0.620***
	(3.6e+05)	(12.273)	(0.056)	(0.056)	(2.1e+04)	(0.164)
Firm algorithm recommends hire	5.3e+05	22.190	0.016	-0.072	6999.1	0.067
	(4.1e+05)	(13.785)	(0.058)	(0.063)	(2.4e+04)	(0.189)
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

Table 6: Performance of employees depending on recommendation by both algorithms. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays and productivity, and marginal effects of probit regressions of the bonus and left-the-firm dummies on dummies for each algorithm's recommendation. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

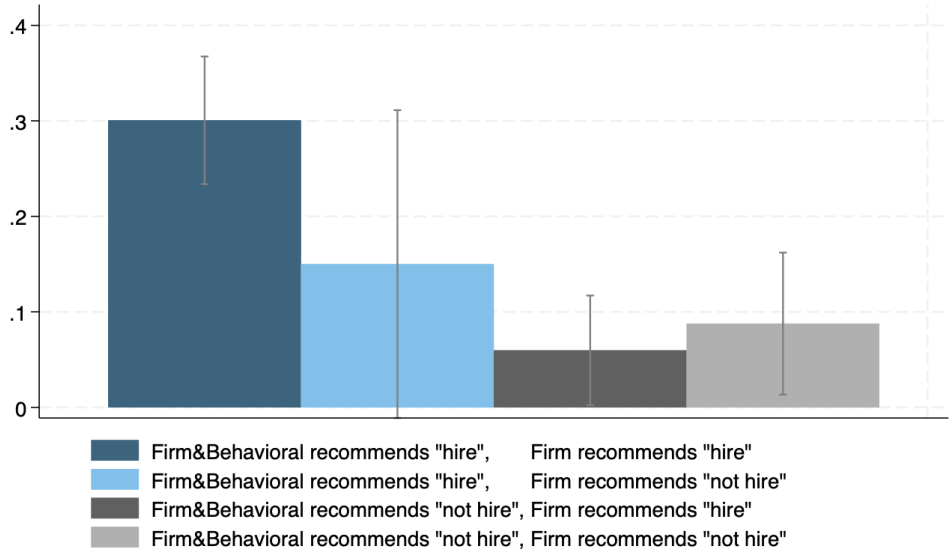

Figure 10: Proportion of employees who obtained a bonus depending on the recommendation of the two algorithms.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

Finally, we compare the profitability of employees identified by each algorithm, using our measure of profitability (see Section 6.2). Note that this is each algorithmic prediction's association with profitability, independent of the other algorithm. Table 7 reports regression results using the linear transformation of 12-month profitability of every employee as the dependent variable. Both algorithms significantly predict profitability, but the effect is larger and more precisely estimated for the algorithm that includes behavioral measures. Interestingly, the constant terms in both models are negative, suggesting that employees not recommended by either algorithm tend to generate losses on average, although these constants are not statistically significant. Based on these estimates, the average profit per new hire is 13.4 percent higher when using the algorithm with behavioral measures compared to the version trained only on firm data.

	Profit for 12 months	Profit for 12 months
Behavioral algorithm recommends hire	8.6e+04*** (3.1e+04)	
Firm algorithm recommends hire		7.2e+04** (3.5e+04)
Constant	-2.1e+04 (2.7e+04)	-1.4e+04 (3.2e+04)
Observations	536	536

Table 7: Profit of employees for 12 months depending on recommendation by both algorithms.
Coefficients of OLS regression for the outcome on an algorithm dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

8 Discussion and conclusions

Our findings offer a positive perspective on the use of behavioral measures in hiring processes, thereby pointing to a new application for behavioral economics: providing standardized measures of human behavior to improve algorithmic hiring and potentially other AI applications, such as credit scoring. In hiring contexts where applicants have few observable qualities to signal their potential and where the job is such that an ideal candidate profile cannot be defined easily, survey measures enable firms to conduct more effective automated screening. It is encouraging that the algorithmic predictions are robust to selective training samples and strategic responses of candidates. While we observe some strategic behavior, particularly for personality measures such as neuroticism, agreeableness, and locus of control, the economic measures show only a small variance between the training sample and the pool of candidates, allowing the algorithm to accurately predict the employees' future performance. One possible concern is that applicants may gradually learn how to answer the survey questions used for algorithmic hiring. Our findings suggest that this concern may be less severe than commonly feared. Although we detect evidence of strategic responses for a subset of survey measures, applicants do not become more likely to be recommended by the algorithm over the one-year duration of the experiment. A likely explanation is that the mapping from survey responses to hiring recommendations is difficult to infer: while some measures exhibit clear directional relationships with hiring recommendations, others enter the prediction in non-monotonic ways, making it difficult for applicants to identify a response profile that systematically improves their predicted score. More generally, our results suggest that the complexity of modern machine-learning algorithms may provide some protection against strategic manipulation, even when applicants know that survey responses are used in hiring decisions. The study was conducted in the specific context of a microcredit firm in Kyrgyzstan. It is uncertain how well our approach performs in other environments. However, the usefulness of behavioral measures and their relative robustness to strategic manipulation can be expected to extend to other recruitment contexts. The tasks of our loan officers involve agency and balance between short-term and longer-term incentives which we believe applies to many white-collar jobs and thus, behavioral features are useful signals of productivity. Although surveys have been utilized by HR for decades, we are unaware of causal evidence that selection based on survey measures outperforms traditional methods, such as interviews. Thus, our results can be interpreted as validating and refining the technique that Daniel Kahneman introduced in the Israeli army in 1955 to improve its hiring practices. He recommended relying on a set of predetermined tests instead of forming intuitive judgments based on interviews. We add the

selection and weighting of these traits with the help of AI and demonstrate its robustness to strategic responses and a biased training sample.

References

Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). (2019). *The economics of artificial intelligence: an agenda*. University of Chicago Press.

Alan, S., Boneva, T., & Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, *134*(3), 1121-1162.

Alan, S., Çorukçuoğlu, G., & Sutter, M. (2023). Improving Workplace Climate in Large Corporations: A Clustered Randomized Intervention. *The Quarterly Journal of Economics*, 138(1), 151–203.

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in OECD countries: A comparative analysis. Working paper.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40-79

Ash, E., Galletta, S., & Giommoni, T. (2020). A Machine Learning Approach to Analyzing Corruption in Local Public Finances. Working paper.

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.

Autor, D. H., & Scarborough, D. (2008). Does job testing harm minority workers? Evidence from retail establishments. *The Quarterly Journal of Economics*, 123(1), 219-277.

Avery, M., Leibbrandt, A. & Vecci, J. (2023). Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech. Working paper.

Awad, E., Balafoutas, L., Chen, L., Ip, E., & Vecci, J. (2023). Artificial Intelligence and Debiasing in Hiring: Impact on Applicant Quality and Gender Diversity. Working paper.

Awuah, K., Krenk, U., & Yanagizawa-Drott, D. (2025). Automation with Generative AI? Evidence from a Teacher Hiring Pipeline. Working paper.

Barsky, R. B., Juster, F. T., Kimball, M. S., & Shapiro, M. D. (1997). Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. The *Quarterly Journal of Economics*, 112(2), 537-579.

Beck, E. D., & Jackson, J. J. (2022). A mega-analysis of personality prediction: Robustness and boundary conditions. *Journal of Personality and Social Psychology*, 122(3), 523.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Harari, Y. N., ... & Mindermann, S. (2023). Managing AI risks in an era of rapid progress. Working paper.

Bhalerao, K., Kumar, A., Kumar, A., & Pujari, P. (2022). A study of barriers and benefits of artificial intelligence adoption in small and medium enterprise. *Academy of Marketing Studies Journal*, *26*, 1-6.

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, 152(1), 4-27.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317-335.

Blader, S., Gartenberg, C., & Prat, A. (2020). The Contingent Effect of Management Practices. *Review of Economic Studies*, 87(2), 721–749.

Bó, I., Chen, L., & Hakimov, R. (2024). Strategic responses to personalized pricing and demand for privacy: An experiment. *Games and Economic Behavior*, 148, 487-516.

Bogen, M., & Rieke, A. (2018): Help Wanted: An Exploration of Hiring Algorithms, Equity and Bias. Working paper.

Bonatti, A., & Cisternas, G. (2020). Consumer scores and price discrimination. *The Review of Economic Studies*, 87(2), 750-791.

Bowles, S., Gintis, H., & Osborne, M. (2001). The determinants of earnings: A behavioral approach. *Journal of Economic Literature*, 39(4), 1137-1176.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, *129*(3), 1409-1447.

Cai, J., & Wang, S. Y. (2022). Improving Management through Worker Evaluations: Evidence from Auto Manufacturing. *The Quarterly Journal of Economics*, 137(4), 2459–2497.

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and selection of human capital with machine learning. *American Economic Review*, *106*(5), 124-127.

Chapman, J., Ortoleva, P., Snowberg, E., Yariv, L., & Camerer, C. (2025). *Reassessing qualitative self-assessments and experimental validation*. Working paper.

Corgnet, B. (2023). An experimental test of algorithmic dismissals. Working paper.

Dastin, J.: Amazon scraps secret AI recruiting tool that showed bias against women. Reuters (2018).https://www.reuters.com/article/us-amazon-com-jobs-automation-insightidUSKCN1MK08G

Dargnies, M. P., Hakimov, R., & Kübler, D. (2026). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*, 72(1), 285-301.

Dean, M. & Ortoleva, P. (2019). The empirical relationship between nonstandard economic behaviors. *Proceedings of the National Academy of Sciences*, 116(33):16262–16267.

Dietvorst, B. J, Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology*: General* 144.1, 114-126.

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2009). Homo reciprocans: Survey evidence on behavioural outcomes. *The Economic Journal*, 119(536), 592-612.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522-550.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, *133*(4), 1645-1692.

Friebel, G., Heinz, M., Hoffman, M., & Zubanov, N. (2023). What do employee referral programs do? Measuring the direct and overall effects of a management practice. *Journal of Political Economy*, 131(3), 633-686.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Gosnell, G. K., List, J. A., & Metcalfe, R. D. (2020). The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains. *Journal of Political Economy*, 128(4), 1195–1233.

Gottfredson, L. S. (2002). g: Highly general and highly practical. In *The general factor of intelligence* (pp. 343-392). Psychology Press.

Guenzel, M., Kogan, S., Niessner, M. & Shue, K. (2026). AI Personality Extraction from Faces: Labor Market Implications. Working paper.

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of machine learning research*, 3(Mar), 1157-1182.

Hackethal, A., Kirchler, M., Laudenbach, C., Razen, M., & Weber, A. (2023). On the role of monetary incentives in risk preference elicitation experiments. *Journal of Risk and Uncertainty*, 66(2), 189-213.

Haeckl, S., & Rege, M. (2024). Effects of Supportive Leadership Behaviors on Employee Satisfaction, Engagement, and Performance: An Experimental Field Investigation. *Management Science*.

Hagenbach, J., & Salas, A. (2025). Strategic information disclosure to classification algorithms: an experiment. *Experimental Economics*, 1-22.

Hakimov, R., Schmacker, R., & Terrier, C. (2023). Confidence and college applications: Evidence from a randomized intervention (No. SP II 2022-209). Working paper.

Hanushek, E. A., Kinne, L., Sancassani, P., & Woessmann, L. (2023). *Can Patience Account for Subnational Differences in Student Achievement? Regional Analysis with Facebook Interests*. National Bureau of Economic Research No. w31690.

Heckman, J. J., Jagelka, T., & Kautz, T. D. (2019). *Some contributions of economics to the study of personality* (No. w26459). National Bureau of Economic Research.

Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology* 1(3), 333-342.

Hoffman, M., Kahn, L. B., & Li, D. (2018). Discretion in hiring. *The Quarterly Journal of Economics*, *133*(2), 765-800.

Hoffman, M., & Stanton, C. T. (2024). People, Practices, and Productivity: A Review of New Advances in Personnel Economics. *Handbook of Labor Economics* 6 (2025): 729-818.

Hossain, T., & List, J. A. (2012). The Behavioralist Visits the Factory: Increasing Productivity Using Simple Framing Manipulations. *Journal of Political Economy*, 120(3), 509–541.

Jabarian, Brian, and Luca Henkel (2025). Voice AI in Firms: A Natural Field Experiment on Automated Job Interviews. Working paper.

Jagelka, T. (2024). Are economists' preferences psychologists' personality traits? A structural approach. *Journal of Political Economy*, 132(3), 910-970.

Kaibel, C., Koch-Bayram, I., Biemann, T., & Mühlenbock, M. (2019). Applicant perceptions of hiring algorithms-uniqueness and discrimination experiences as moderators. In *Academy of Management Proceedings* (Vol. 2019, No. 1, p. 18172).

Kaur, S., Kremer, M., & Mullainathan, S. (2015). Self-Control at Work. *Journal of Political Economy*, 123(6), 1227–1277.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. Working paper.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237-293.

Kohavi, R. "A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection." *Ijcai*. Vol. 14. No. 2. 1995.

Komarraju, M., Karau, S. J., Schmeck, R. R., & Avdic, A. (2011). The Big Five personality traits, learning styles, and academic achievement. *Personality and individual differences*, 51(4), 472-477.

Krueger, M., & Friebel, G. (2022). A pay change and its long-term consequences. *Journal of Labor Economics*, 40(3), 543-572.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 2053951718756684.

Li, D., L. Raymond and P. Bergman (2026). Hiring as Exploration. *Review of Economic Studies* 93.2 (2026): 1200-1240.

Lönnqvist, J. E., Verkasalo, M., Walkowitz, G., & Wichardt, P. C. (2015). Measuring individual risk attitudes in the lab: Task or ask? An empirical comparison. *Journal of Economic Behavior & Organization*, 119, 254-266.

Mammadov, S. (2022). Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality*, *90*(2), 222-255.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H. & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89-94.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel psychology*, 60(3), 683-729.

Mullainathan, S., & Obermeyer, Z. (2022). Diagnosing physician error: A machine learning approach to low-value health care. *The Quarterly Journal of Economics*, 137(2), 679-727.

Radhakrishnan, J., & Chattopadhyay, M. (2020). Determinants and barriers of artificial intelligence adoption–A literature review. In *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation: IFIP WG 8.6 International Conference on Transfer and Diffusion of IT, TDIT 2020, Tiruchirappalli, India, December 18–19, 2020, Proceedings, Part I* (pp. 89-99). Springer International Publishing.

Rhea, A.K., Markey, K., D'Arinzo, L. *et al*. (2022). An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Min Knowl Disc* 36, 2153–2193.

Roulin, N., & Krings, F. (2020). Faking to fit in: Applicants' response strategies to match organizational culture. *Journal of Applied Psychology*, *105*(2), 130.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. *The Stata Journal*, 20(1), 3-29.

Vieider, F. M., Lefebvre, M., Bouchouicha, R., Chmura, T., Hakimov, R., Krawczyk, M., & Martinsson, P. (2015). Common components of risk and uncertainty attitudes across contexts and domains: Evidence from 30 countries. *Journal of the European Economic Association*, 13(3), 421-452.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and psychological measurement*, *59*(2), 197-210.

Zhang, S., & Kuhn, P. J. (2024). Measuring Bias in Job Recommender Systems: Auditing the Algorithms. Working paper.