RATIONALITY & COMPETITION
CRC TRR 190

Behavioral Measures Improve AI Hiring: A Field Experiment

Marie-Pierre Dargnies (University of Paris Dauphine)
Rustamdjan Hakimov (University of Lausanne)
Dorothea Kübler (WZB Berlin, Technische Universität Berlin, CES Ifo)

Discussion Paper No. 532

April 29, 2025

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Dr. Klaus M. Schmidt, Ludwig Maximilian University of Munich, 80539 Munich, Germany
info@rationality-and-competition.de

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

Marie-Pierre Dargnies (University of Paris Dauphine, PSL)
Rustamdjan Hakimov (University of Lausanne)
Dorothea Kübler (WZB Berlin, Technische Universität Berlin & CESIfo)

April 2025

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. 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. The comparison of workers hired by the algorithm with those hired by human managers in the field experiment reveals that algorithmic hiring is marginally more efficient than managerial hiring.

Keywords: Hiring; AI; economic and behavioral measures; selective labels

[1] We thank Brice Corgnet, Guido Friebel, Johannes Johnen, Amma Panin, Christian Zehnder, and participants of 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, seminar participants at Utrecht University, ECARES Brussels, University of Southern Denmark, 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_207722). Dorothea Kübler acknowledges financial support by the Deutsche Forschungsgemeinschaft through CRC TRR190 ("Rationality and Competition").

1 Introduction

Artificial intelligence and more specifically machine learning have made significant progress lately (Goodfellow et al. 2016, Bengio et al. 2023). While automation had an impact mainly on jobs with a high proportion of routine tasks (Autor 2015, Arntz et al. 2016), many high-skilled jobs will likely be affected by the rise of machine learning. Examples of potential replacements abound. AI is found to outperform human decision-making in screenings of breast cancer (McKinney et al. 2020) and diagnosing heart attacks (Mullainathan and Obermeyer 2022), jail or release decisions (Kleinberg et al. 2018), the detection of corruption (Ash et al. 2020), as well as police hiring and teacher tenure decisions (Chalfin et al. 2016).

At the same time, there are obstacles to the adoption of artificial intelligence by firms and organizations (Agrawal et al, 2019, Radhakrishnan et al, 2020). Managers, clients, and workers can be averse to algorithms (Highhouse 2008, Dietvorst et al. 2015, Dargnies et al. 2024), e.g., because of potential biases of AI-driven hiring compared to traditional methods. Also, the use of AI requires expertise and the availability of large data sets which small firms may lack (Bhalerao et al., 2022). In response, more and more firms specialize in AI-based evaluations of job candidates as a service, screening online applications and CVs, and this market is projected to grow rapidly.[2] However, these algorithms are typically not developed and trained on data of the client firms, casting doubt on their effectiveness.

In this paper we study whether behavioral data collected via an employee survey can be used as an input to an AI-trained algorithm predicting the productivity of loan officers in a microfinance company. In the microfinance sector, hiring prerequisites are minimal, and the firm we are collaborating with collects only a small set of employee characteristics. At the same time, productivity varies significantly among employees, with particularly high turnover rates observed among those with lower productivity levels. Thus, additional measures could be helpful for scoring the applicants. We concentrate on cognitive and behavioral measures that are widely used in economics and psychology. These measures of behavioral traits and preferences are known to correlate with education, employment, and investment decisions as well as related outcomes (Barsky et al. 1997, Dohmen et al. 2011, Gottfredson 2002, Dohmen et al. 2009, Buser

[2] https://www.maximizemarketresearch.com/market-report/global-ai-recruitment-market/63261/?utm_source=chatgpt.com, last accessed on 20.02.2024.

et al. 2014, Hakimov et al. 2023, Alan et al. 2019, Hanushek et al. 2023). Also, research in psychology has advocated the use of personality tests to predict job performance, though the evidence remains mixed and has been the subject of debates (Morgeson et al. 2007).

To assess whether a prediction algorithm based on behavioral measures can improve hiring decisions, all loan officers at the microfinance firm were required to complete a survey. This survey included both incentivized and non-incentivized behavioral measures, eliciting risk and time preferences, trust, trustworthiness, and altruism, alongside measures from psychology and cognitive tests. We trained random-forest algorithms on data from employees with at least 12 months of tenure, a time period that was deemed sufficient for a reliable productivity assessment by the firm's managers. We employed a pre-registered binary productivity metric as defined by the management of the firm: whether employees qualified for a bonus in addition to their salary within the first year of employment. This measure is objective and is based on the size of the portfolio managed by each loan officers, as well as the quality of the repayments.

A random forest algorithm, relying solely on firm data such as age, education level, marital status, and other factors, accurately categorized 65% of employees in an out-of-sample test. The inclusion of the non-incentivized measures and tests 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. For the next steps of the project, we therefore apply the random forest algorithm trained on firm data and non-incentivized measures which are easier to elicit for firms than incentivized measures.

There are two potential concerns regarding the applicability of AI trained with survey measures for hiring decisions. One concern stems from the fact that our training sample is the result of past hiring decisions made by HR as well as decisions of employees to accept the job and stay with the company for at least one year, making it susceptible to what the literature identifies as the *selective labels* problem (Kleinberg et al. 2018, Chalfin et al. 2016). This selection could potentially lead to an inaccurate weighting of key characteristics in the algorithm when applied to a sample of new applicants. The second concern is that applicants may strategically alter their answers to align with their perception of the company's ideal employee profile, unlike existing

employees whose data trained the random forest algorithm and had no incentive to manipulate their responses. This may also impact the algorithm's performance.

As a second step, after having trained and selected an algorithm that uses some of the firm data and behavioral measures, we address the selective labels 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. We find that those predicted to be productive by our algorithm are significantly more likely to receive a bonus both at the 12-month tenure mark and one year following the survey, both conditional and unconditional on being employed when the bonus decision is made. They are also significantly less likely to leave the firm within the first year of employment and within one year after the survey, consistent with positive selection in the workplace.

In the third stage of our study, we conducted a field experiment on hiring new employees with two primary goals. First, it allows us to further address the selective labels problem by predicting the performance of newly hired employees and assess whether potential manipulations of survey responses by applicants affect the algorithm's accuracy. Second, we can evaluate the effectiveness of algorithmic hiring compared to the firm's current hiring practices. 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.

We observe that candidates recommended for hiring by the algorithm were significantly more likely to receive a bonus at the 12-month mark and in February 2024 (one year after the end of the experiment), compared to those not recommended. However, there was a significant difference in retention rates between the two groups only in February 2024, but not on the 12-month mark of employment. Based on these findings, we conclude that the algorithm is robust to both selection effects and the strategic manipulation of responses, although its predictive power regarding retention is weaker in the applicant sample than in the sample of employees. We investigated to what extent this can be attributed to strategic responses. A comparison of the distribution of responses to the survey between existing employees and candidates reveals several

significant differences. 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.[3] Another way to identify strategic responses is to analyze the proportion of applicants recommended for hiring by the algorithm over time. If gaming occurs, this proportion should rise, but we find no evidence of such a trend. 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.

To address the second goal of the experiment—assessing the effectiveness of algorithmic hiring relative to the firm's current HR practices—we compared employee performance under the two experimental treatments at the 12-month mark, and in February 2024 when the last hires of the experiment reached 12 months of employment. During the experiment, the managers rejected fewer candidates than expected, reducing variation across treatments.[4] Nonetheless, employees hired in the AI treatment performed better overall than those hired in the HR treatment, though differences are only marginally significant without considering employment status (i.e., counting those who left as not obtaining a bonus). When considering only those who are still employed, the likelihood of receiving a bonus at the 12-month mark and in February 2024 is significantly higher in the AI treatment, suggesting the algorithm outperformed the HR selection process.

Overall, our paper provides a constructive framework to train 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. The success of this approach heavily depends on whether certain traits are universally desirable and whether resumes provide sufficient information to evaluate the productivity for a certain job. Our results shed doubt on this by demonstrating the value of behavioral measures.

[3]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).

[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

Related literature

This study contributes to different strands of the literature. Personality traits and preferences have been shown to predict and cause important outcomes such as wages, health, and longevity (Heckman et al. 2019). Existing work has mainly focused on the correlation or the causal effect of one of such measures on economic decisions or outcomes. For example, risk preferences are correlated with portfolio choice and self-employment (Dohmen et al. 2011), answers to the Big-5 test predict sorting into careers and academic performance (Gottfredson 2002), and positive reciprocity is associated with higher earnings (Dohmen et al. 2009). Competitiveness and confidence are positively correlated with the choice of more prestigious school-tracks and causally affect university choices (Buser et al. 2014, Hakimov et al. 2023), while grit and patience lead to higher educational attainments (Alan et al. 2019; Hanushek et al. 2023). Economists have also documented that behavioral measures are important determinants of lifetime earnings (see Bowles et al. 2001 and Kautz et al. 2014 for extensive surveys). Similarly, research in psychology shows that personality traits are associated with life outcomes (see Beck and Jackson 2022 for a review and meta-study of the robustness of these findings). While associations are well established for many measures, the predictive power of each of them is low and often not the focus of the work. We investigate whether a combination of the measures can be put to work to improve the prediction of employee performance. Moreover, we run a field experiment where hiring decisions are based on behavioral measures, thereby testing their robustness in a setting where applicants may provide strategic responses.

The relationship between psychological and behavioral measures is studied by Dean et al. (2019) and Jagelka (2024). We contribute to this research by testing the effectiveness of various combinations of measures in predicting productivity. Our findings demonstrate that economic and psychological measures can complement each other. We also provide a novel method to investigate and compare the measures' robustness to strategic manipulation. There is suggestive evidence that economic measures are more robust to manipulations than the Big Five personality traits.

Our paper also contributes to the literature exploring how tests, machine learning, and AI can contribute to HR practices. Hoffman and Stanton (2024) provide a comprehensive review of recent work, including a section on the impact of technology and other procedures on hiring practices. Autor and Scarborough (2008) show that job testing improves selection but reduces minority hiring. Hoffman et al (2018) find that managers who hire against test recommendations

select applicants with lower subsequent retention rates. A number of papers consider the efficiency of AI tools for hiring and tenure decisions. Chalfin et al. (2016) demonstrate the superior performance of AI relative to human decisions in a data-rich environment, namely public sector hiring and tenure decisions. In contrast to their work, we address the so-called 'selective labels' problem by conducting a field experiment. Four studies consider the effect of AI hiring on labor demand and supply. Avery et al. (2023) employ a field experiment to study labor demand and supply. They find a positive effect of AI hiring on the number of women's applications in a stereotypically male profession, and more favorable evaluations of women when managers receive AI scores of candidates than without them. Second, Awad et al. (2023) employ simulated labor market experiments to explore the impact of AI utilization and the debiasing of both humans and AI on the quality and gender diversity of job applicants. Their findings reveal that the use of AI does not alter the quality and gender diversity of applicants when compared to assessments by human evaluators. Debiasing humans or AI improves gender diversity without reducing the number of high-quality applicants. Third, in an audit study of a job recommender algorithm, Zhang and Kuhn (2024) find that otherwise identical male and female applicants do not receive the same job recommendations. Moreover, the jobs recommended to men and women exhibit different characteristics, particularly in terms of wages. Finally, Li et al. (forthcoming) compare an exploratory to a traditional supervised learning algorithm and find that the quality of applicants selected by the exploratory algorithm is better, and more good candidates from underrepresented groups are hired.

Some authors have studied the general acceptance of the use of AI for hiring decisions, such as Lee (2018), Kaibel et al. (2019), Bigman et al. (2022) and Corgnet (2023). The first two articles show that people have reservations against AI hiring while the last two discover more negative reactions to human than to algorithmic decisions.

A number of recent studies have explored the capacity of agents to adapt their behavior strategically in reaction to profiling efforts. Bonatti and Cisternas (2020) provide a theoretical examination of the consequences of consolidating consumers' purchase histories into proxies for unobserved willingness to pay. They find that the welfare implications critically depend on the consumers' strategies to manipulate their proxies. Bo et al. (2023) demonstrate experimentally that when participants are aware that the price of lottery tickets may be personalized, they can successfully alter their responses to surveys related to risk measures, thereby lowering prices. Similarly, Hagenbach and Salas (2024) show that the majority of participants in a lab experiment

fails to optimally conceal their answers from the algorithm, thus allowing for profiling. In a meta-analysis, Viswesvaran and Ones (1999) compared fake and honest responses to personality measures classified under the Big Five dimensions. Their findings reveal that participants instructed to fake positive responses scored higher across all Big Five dimensions compared to those instructed to respond honestly. In the hiring domain, Birkeland et al. (2006) conducted a meta-analysis comparing personality scale scores between job applicants and non-applicants, revealing significant differences in extraversion, conscientiousness, and openness. However, they acknowledge that these differences may be influenced by selection effects. Roulin and Krings (2020) demonstrate in experimental and survey studies that participants in the role of applicants tailor their personality profiles—specifically in terms of competitiveness and innovativeness—based on what they perceive to be the ideal fit for the organization's culture. Our research contributes to this body of literature by investigating the strategic responses of job applicants in the survey-based screening process of the AI and by going beyond psychological traits and including economic measures. We examine the differences in responses between job applicants and employees, controlling for potential differences between cohorts and selection effects.

Our paper adds to the broader literature that uses insights from behavioral economics to improve firm policies. Hossain and List (2012) find that framing bonuses as losses instead of gains increased productivity in a Chinese factory, showing the power of framing in motivating employees. Gosnell et al. (2020) observe that personalized feedback and goal-setting improved fuel efficiency among airline captains, demonstrating the effectiveness of behavioral nudges in high-pressure environments. Offering commitment devices to employees reduces procrastination and increases productivity, see Kaur et al. (2015). Interventions targeting social norms and communication improve workplace climate, raising employee satisfaction and engagement, as shown by Alan et al. (2023). Blader et al. (2020) find that management practices work best when they align with employees' perceptions and intrinsic motivations, and Cai and Wang (2022) show that including worker evaluations in management decisions boosted productivity in auto manufacturing. Krueger and Friebel (2022) demonstrate that fairness and reference points are important constraints in policies related to payment scheme reforms. Friebel et al. (2023) show that the effectiveness of employee referral programs is driven by employees valuing their involvement in the hiring process. Our study contributes to this strand of the literature by demonstrating that behavioral measures can improve hiring by making AI-driven processes more accurate, especially when data collected by the firm are limited.

Finally, our study speaks to the behavioral economics literature discussing relations between incentivized and non-incentivized measures. In a representative German sample, Dohmen et al. (2011) compare risk-attitude measures obtained via survey answers with incentivized responses to a 50-50 prospect, and found a significant correlation between the two. Vieider et al. (2013) replicate this result in a large sample of 30 countries. Lönnqvist et al. (2011) assess the risk attitude survey question's intertemporal stability in comparison to an incentivized task developed by Holt et al. (2002). They find that the survey question performed better. Falk et al. (2018) explore non-incentivized measures for various economic behaviors, such as time preference, trust, and others. Hackethal et al. (2023) have conducted a series of experiments comparing incentivized and non-incentivized risk elicitation methods and conclude that incentives do not significantly affect the results. Our results contribute to this literature by showing that survey measures can serve as substitutes for incentivized measures, although this requires a larger number of questions.

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 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.

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 employees took the survey. The average duration was one hour and six minutes. There were 1042 employees among which 674 were employed by the firm for 12 months or more. Following the management's assessment that one year is typically enough to qualify for the company bonus, we used these 674 employees to train the algorithm.
The survey was both 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 key 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. The reason for restricting the sample to this group is that their productivity is more reliably observed than for employees with a shorter history at the firm. We train the AI, a random-forest algorithm, to classify employees according to the pre-registered binary variable of being a high-achieving employee or not, defined by the reception of a bonus within one year of employment.[5] Algorithms predicting other variables (longer-term performance, turnover, portfolio at risk, size of portfolio) are run for exploratory purposes.

For the prediction of our main outcome variable, i.e., the payment of a bonus, one algorithm uses only personnel data (such as gender and age), another algorithm uses both the personnel data and the answers to the non-incentivized questions, and a third algorithm uses the personnel data, the answers to the non-incentivized questions, and the answers 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 the 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 algorithms' predictions. This allows us to measure the out-of-sample efficiency of the algorithms. 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 goal of the field experiment is twofold. First, it enables us to test the robustness of the algorithm to sample selection and potential strategic responses from candidates. Second, it allows us to evaluate the firm's current HR practices against algorithmic hiring. For the experiment, some employees were hired following the normal procedure of the firm (applicants are interviewed by the local office

[5] Bonus payments vary in size, depending on the profitability of the employee's portfolio. For successful employees they reach around $3000, while the base salary is only around $200.

manager and by a senior manager, i.e., the manager of a region) while others were hired following the recommendation of the AI algorithm.[6] 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 the 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 (the manager 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. 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 transmitted 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.[7] For the other half of the applicants, i.e., those in the HR treatment, however, senior management followed their usual decision process without receiving the recommendation of the algorithm. Note that due to this design, we know for every applicant the hiring recommendation by HR and by the algorithm.

Neither the applicants nor the local managers were aware of their participation in a study. Everyone knew about the survey in the first stage, but they were not informed of its role for the

[6] 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.
[7] 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.

algorithm's recommendations or that this recommendation was followed for some applicants but not for others. Only the board knew about the experiment involving the algorithm. Applicants were merely 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 (DOI:10.1257/rct.8219-1.0).

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 entropy as the splitting criterion. We calibrated the parameters of the algorithm to reach the highest performance, where the performance was evaluated based on the out-of-sample prediction accuracy in the following way: We ran the algorithm 250 times, based on a random split of the data into 550 observations for training and 124 observations for validation. We recorded the percentage of correctly classified employees in the validation sample for each random split. Based on calibrations, 10 variables were randomly selected for each split, and 500 sub-trees were constructed at maximum.

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 select the variables for the final algorithm, we began with a comprehensive algorithm encompassing all collected variables. We then systematically removed the least important variables for tree splitting, continuing until the algorithm's average accuracy declined with further variable removal.

Based on these procedures and only using the final set of non-incentivized behavioral measures, we can study a number of questions regarding the performance of the algorithm. For example, how much does the algorithm improve when using non-incentivized behavioral and

psychological traits in addition to firm data alone? How much worse does the algorithm get when not using the incentivized measures?

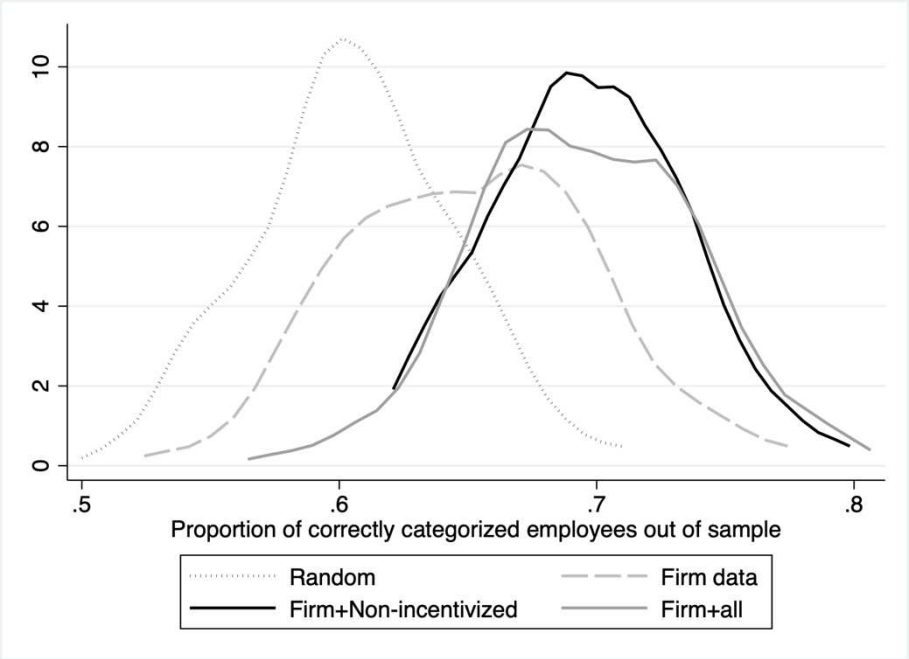


Figure 1: Proportion of employees correctly categorized by the Random-Forest Algorithm depending on set of variables used.
Notes: Only the final set of non-incentivized measures is used, based on the selection process described above. The number of trees, iterations, and the sample size per tree are calibrated for each algorithm separately.

Figure 1 presents the distribution of accuracy of random forest algorithms depending on the set of explanatory variables used. On average, the random forest algorithm using only firm data classifies 65.1% of the 124 loan officers in the validation sample correctly. This is significantly better than an algorithm based on ten randomly generated variables (a two-sided Fisher's exact test yields $p<0.001$). Including the non-incentivized survey measures significantly improves the algorithm's performance 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 improve the performance further ($p=0.48$).[8] The firms' use of 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.

[8] 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.

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 the algorithm does better than the probit model when firm data and non-incentivized measures are used ($p<0.001$). The magnitude of the effect is 1.5 percentage points, which is one-third of the improvement in accuracy from using the non-incentivized measures. This improvement of the random forest vis-á-vis a probit regression can be due to the effects of the non-incentivized measures being either non-linear or resulting from interactions between different variables.

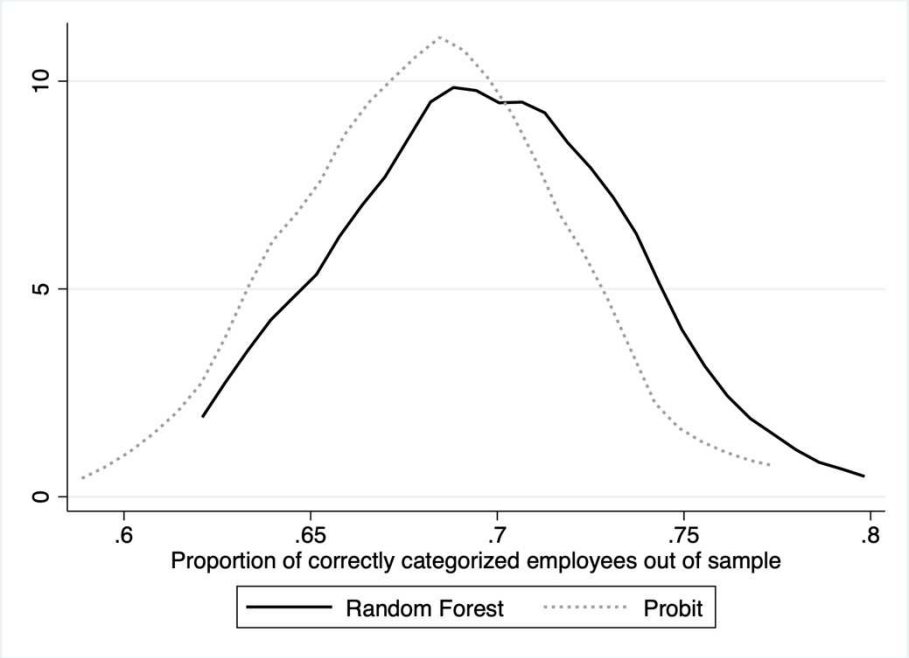


Figure 2: Proportion of correctly categorized employees for the random forest algorithm and probit regression.
Notes: Only the final set of non-incentivized measures is used, based on the selection process described above.

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 the algorithm using only firm data is significantly improved by adding gender and ethnicity ($p<0.001$), while this is not the case for the algorithm that also uses the non-incentivized survey measures ($p=0.20$).[9]

[9] 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 only (average out-of-sample accuracy 59.5%), and the other based on the non-incentivized measures only (average

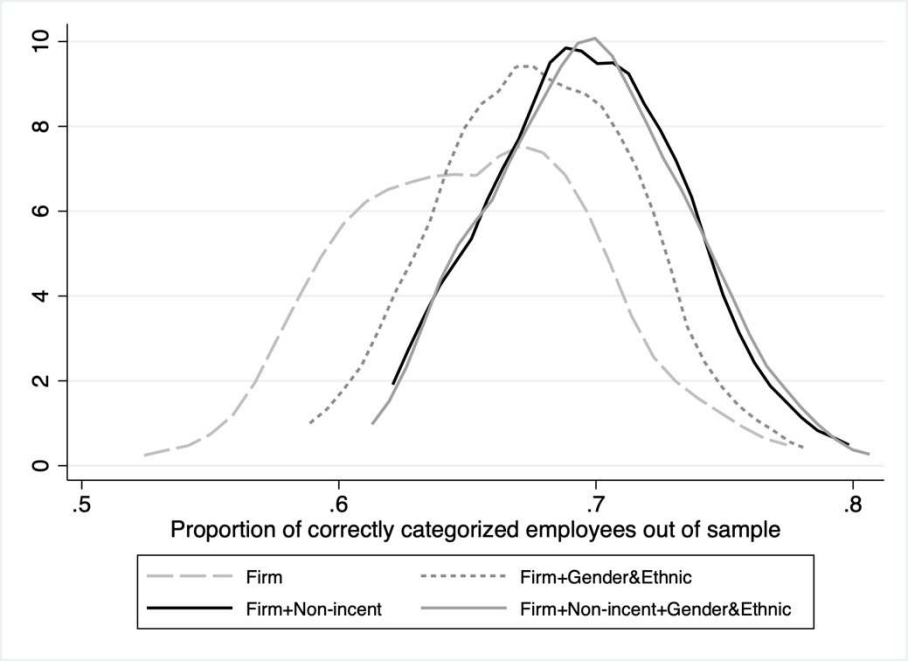


Figure 3: Proportion of correctly categorized employees for algorithms with and without gender and ethnicity.

Based on the performance of the algorithms and the impracticality of using incentivized measures, we select the algorithm based on firm data and a subset of the non-incentivized measures. Table 1 presents the variables retained in the final algorithm. The variables are ordered by importance, from highest to lowest. Importance is a relative measure, where the most frequently used variable for tree splitting is assigned a value of 100%, with the other variables receiving weights relative to this variable. We ran the algorithm 250 times, based on different random splits between the training and the validation sample. Table 1 shows the average importance, calculated from 250 iterations run on distinct subsets of the training data.

The self-reported level of risk aversion on a scale from 1 to 10 emerges as the most important variable, followed by self-reported patience on a scale from 1 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 validated by behavioral economists. The algorithm also uses some firm data, including regional effects represented by dummies for regions managed by different regional managers,[10] age, a dummy

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.

[10]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

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 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 linear.

Variable	**Importance**
Risk aversion 1 to 10	100%
Patience 1 to 10	91.5%
Guess of 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 degree	67.9%
Positive reciprocity 1 to 10	67.7%
Locus of control Rotter scale	67.6%
Agreeableness	65.6%
Specialization Econ or Management	65.2%
Altruism 1 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 aversion 1 to 10 is the response to "How willing or unwilling you are to take risks?"; Patience 1 to 10 is the response to "Would you describe yourself as a patient person?"; "Guess of 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;

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.

"Altruism 1 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.

In Appendix B, we present additional metrics comparing the performance of algorithms based on firm data alone versus firm data combined with non-incentivized measures. Confusion matrices and the density of model-predicted probabilities conditional on receiving a bonus confirm that non-incentivized survey measures robustly improve the algorithm's performance. In the next sections, we investigate the robustness of the algorithm combining firm and non-incentivized data for the selected training sample and candidates' strategic responses.

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, the application of the algorithm 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. We do not have data for those who were not hired or who left the company before completing one year of tenure. This could diminish the algorithm's performance for an unselected sample of candidates if the algorithm is intended to be applied to make hiring decisions. The second concern is that job candidates may answer some survey questions strategically in an attempt to increase their chances of being hired, while this was not a concern for the questions answered in September 2021 by the employees of the firm.

The two concerns are addressed in two steps. In this section, we reduce the problem of selection effects by examining the future performance of employees with less than 12 months of tenure as of September 2021. We assume that their responses were not strategic, or at least not more strategic than those in the training sample, since they were all employed when completing the survey. However, 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. Nevertheless, the sample 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 not.[11]

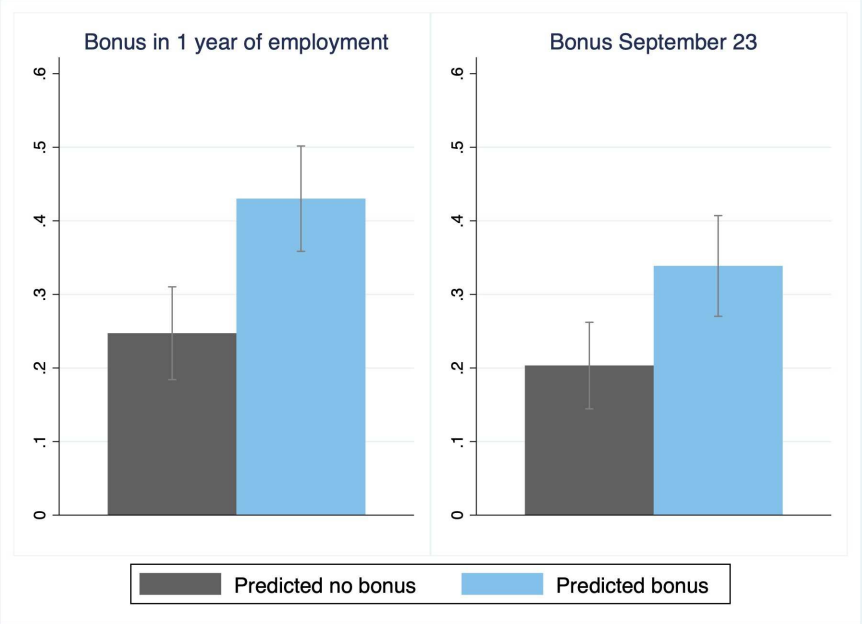


Figure 4: Proportion of employees who obtained a bonus depending on the prediction of the algorithm.
Notes: Gray lines represent 95% confidence intervals. Sample of employees at the firm for < 12 months when responding to the survey.

Figure 4 displays the proportions of employees who actually obtained a bonus, separately for those predicted by the algorithm to get it or not. The left panel of Figure 4 shows the proportion of employees with a bonus after the 12th month of employment, while the right panel shows the proportion of employees with a bonus in September 2023.[12] In both cases, employees who were predicted to get a bonus were significantly more likely to receive it than those who were predicted not to get it (Fisher's exact test $p<0.01$ for both comparisons).

The firm is characterized by a high rate of turnover: more than 50% of new employees leave the firm within one year. Of the 368 employees who were at the firm for less than 12 months in

[11]Note that in the training sample of employees with at least 12 months of tenure, the algorithm predicts 69% of employees to receive a bonus, compared to 50.5% for employees with less than 12 months of tenure. This may be driven by positive selection at work.
[12]We pre-registered the measure of a bonus payment at 12 months of employment, but we also present results for the bonus in September 2023 for exploratory purposes. This second measure allows us to observe the longer-term outcomes for part of the sample. September 2023 is the most recent date for which we have data on employees who had less than one year of tenure as of the original survey in September 2021.

September 2021, 175 had left the firm before reaching the one-year mark and 231 had left the firm by September 2023. The departures were more numerous among those predicted not to get a bonus than among those predicted to get it. Respectively, 39% and 56% of employees predicted to get or not get a bonus did not reach the one-year mark (Fisher exact test $p<0.01$). Respectively 56% and 67% of employees predicted to get a bonus and not to get a bonus had left the company by September 2023 (Fisher exact test $p=0.03$). Note that we cannot distinguish between employees who left the company due to poor performance (or who were fired) and those who left because they found a more attractive job. However, we can observe whether the employees who had left the company by September 2023 received a bonus after one year of employment. Among the employees who left, 17% received a bonus within one year of employment, compared to 62% for those who were still employed in September 2023 (Fisher exact test, $p<0.01$), indicating a strong correlation between receiving a bonus and tenure. Moreover, among those who left, the proportion of employees receiving a bonus was significantly higher for those predicted by the algorithm to receive a bonus (26%) than for those predicted not to receive a bonus (10%) (Fisher exact test, p-value<0.01).

Considering only those employees who were still with the firm at the time of the bonus decision, we look at the proportions of employees who obtained a bonus, depending on whether they were predicted to receive it or not. As can be taken from Figure 5, the employees who were predicted to get a bonus were more likely to receive it than those who were predicted not to receive it (Fisher exact test $p=0.047$ and $p=0.051$, respectively).

Table 2 displays the results of regressions where the dependent variables are obtaining a bonus or leaving the company, with an independent variable indicating whether the algorithm predicts a bonus. Employees predicted to get a bonus are significantly more likely to get it. This result holds when restricting the sample to employees still working at the firm at the time of the bonus decision, excluding those who left the firm from the group of non-recipients. Finally, the regressions show that the algorithm's prediction that an employee will get a bonus is associated with a lower probability of leaving the firm.

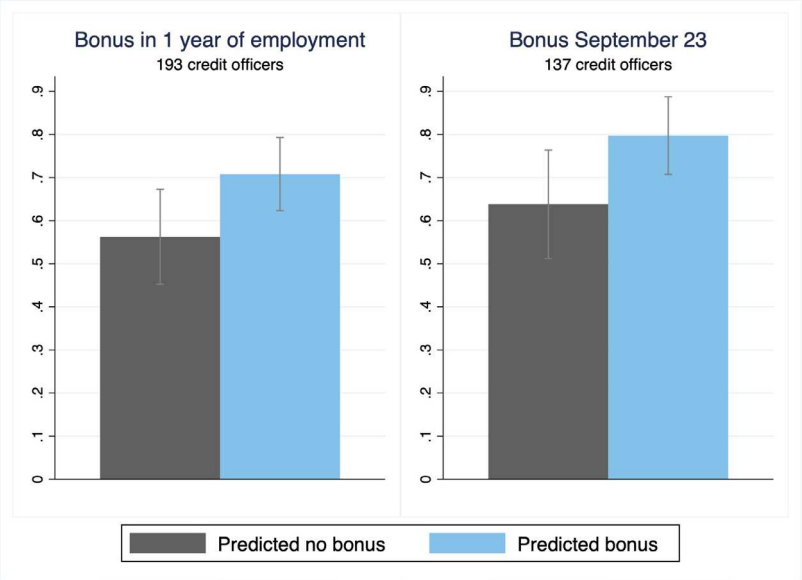


Figure 5: Proportion of employees with the firm who obtained a bonus depending on the prediction of the algorithm.
Notes: Gray lines represent 95% confidence intervals. Sample of employees at the firm for < 12 months when responding to the survey.

	Bonus 1 year	Bonus 09.23	Left 1 year	Left 09.23	Bonus 1 year	Bonus 09.23
Algorithm predicts bonus	0.180*** (0.053)	0.134*** (0.045)	-0.165*** (0.058)	-0.105** (0.051)	0.142** (0.070)	0.154** (0.066)
Observations	368	368	368	368	193	137
Clusters	98	98	98	98	90	79
Sample	All	All	All	All	Employed	Employed

Table 2: Performance and turnover of employees with tenure < 12 months in 1 year and in September 2023. Marginal effect of probit regression of the outcome on the dummy that the algorithm predicts a bonus, with standard errors clustered at the level of offices. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

A field experiment was conducted to evaluate the effectiveness of algorithmic hiring compared to current HR practices. In this section, we present the experimental design. In Section 6, we describe and discuss the results concerning the AI prediction's robustness to sample selection and strategic responses, and in Section 7, we examine the treatment differences to compare algorithmic hiring with current hiring practices by HR.

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.

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. For this calibration of the AI, we used the data from existing employees (the training sample and the sample of employees with short tenure) to determine the threshold for rejections.[13] For hiring, we chose an approach similar to the training of the algorithm, where 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

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

bonus. We used a cutoff of 25, calibrated 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 the 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 to hire 63.7% and 62.6% of applicants in treatments AI and HR, respectively (p=0.72).[14] The managers recommended to hire 96.6% and 97.8% in treatments AI and HR, respectively (p=0.22). Thus, the treatments were balanced with respect to the 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 for most of 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 sample being applicants recruited in the HR treatment who would have been rejected by the AI.

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.[15] In the HR treatment, the share of applicants who were not hired despite the offer was not significantly different depending on the recommendation of the algorithm (43% among those recommended and 44% among those not recommended by the algorithm, p=0.86). 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 Predicting the performance of job applicants

[14] 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.
[15] 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.

There are two concerns regarding the usefulness of our hiring algorithm, the selective training sample and strategic answers to the survey by job applicants. We check for the relevance of these potential biases by running the algorithm on new applicants to the firm and comparing the AI prediction to their actual performance.

6.1 Algorithm predicting the performance of applicants

We first investigate whether employees whom the algorithm recommended to hire are more likely to receive a bonus than those whom the algorithm recommended not to hire. In total, 956 applicants received an offer, which corresponds to 81% of applicants. Of those, 44% (420 applicants) ended up not joining the firm, leaving us with 536 applicants hired of whom 62% were recommended to be hired by the algorithm. Figure 6 shows that the new employees recommended to be hired by the AI more often receive a bonus on the 12th month of employment and in February 2024 than those not recommended to be hired by the AI (the two-sided Fisher exact tests both yield $p<0.01$). This result also holds when restricting the sample to workers still employed when the bonus decision was taken (see Figure 7).[16]

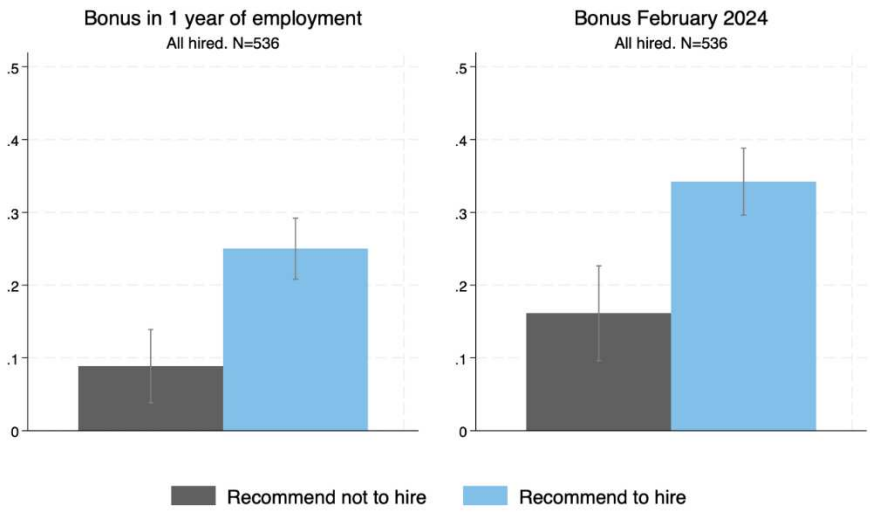


Figure 6: Proportion of employees who obtained a bonus depending on the recommendation of the AI.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

[16] In February 2024, those hired at the end of the experiment had reached 12 months of employment.

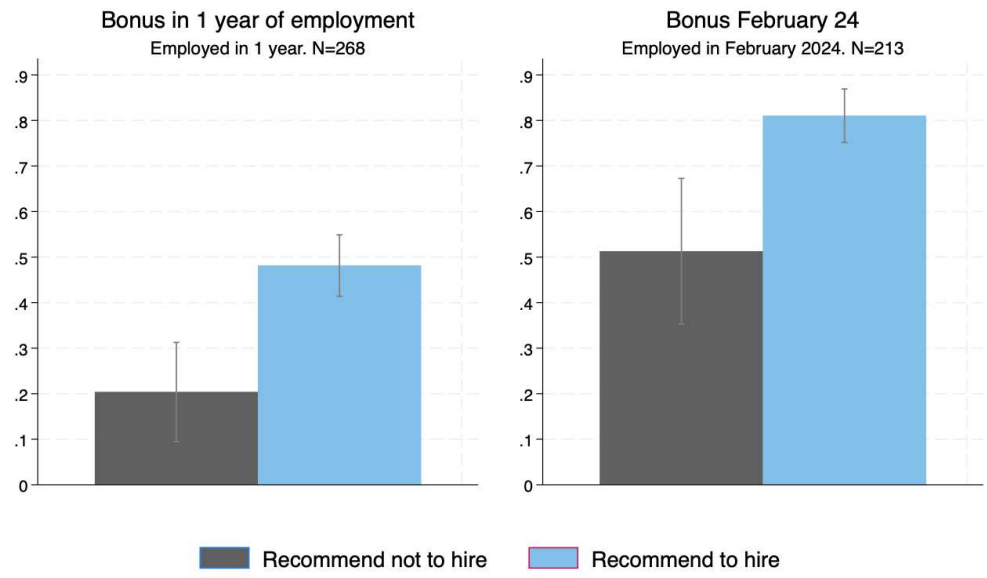


Figure 7: Proportion of employees still in the firm who obtained a bonus depending on the recommendation of the AI.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

	(1)	(2)	(3)	(4)	(5)	(6)
	Bonus 1 year	Bonus 02.24	Left 1 year	Left 02.24	Bonus 1 year	Bonus 02.24
Algorithm recommended hire	0.190*** (0.045)	0.198*** (0.045)	-0.084 (0.056)	-0.112** (0.049)	0.293*** (0.074)	0.249*** (0.065)
Observations	536	536	536	536	268	213
Clusters	102	102	102	102	91	90
Sample	All	All	All	All	Employed	Employed

Table 3: Performance and turnover of employees. Marginal effects of probit regression of the outcome on a dummy for the algorithm recommending hiring, with standard errors clustered at the level of offices. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The probit regressions displayed in Table 3 demonstrate the predictive power of the AI's hiring recommendation for receiving a bonus. Applicants recommended by the algorithm are significantly more likely to receive a bonus both unconditionally and conditional on still being employed after one year or in February 2024. The regression in column 3 shows that leaving the firm within the first year of employment does not occur significantly less often for employees whom the AI recommended hiring. However, those recommended by the algorithm are significantly less likely to leave the company before February 2024 (see the fourth column).

Thus, the random forest algorithm trained on firm data and non-incentivized measures predicts the future performance of the applicants well. The results are similar to the prediction of the algorithm for existing employees with short tenure, except for the lack of a significant difference

in the probability of leaving the firm within one year. We conclude that the predictions of the algorithm regarding bonus payments are robust to the selective training sample and to the potentially strategic responses of the applicants. Thus, the study demonstrates the usefulness of non-incentivized measures as an input for applicant screening.

However, there are some shortcomings to this analysis. Since our data comes from a field experiment, we only observe the performance of about half of the candidates rejected by the algorithm, namely those who were in the HR treatment and were accepted by the managers. This reduction of statistical power is a cost we had to pay to evaluate the effectiveness of algorithmic hiring (see the results in Section 7). Given the near-universal acceptance rate in the HR treatment, sample bias due to the experimental design is limited.[17] The largest potential bias is due to the fact that a large share of candidates (44%) declined the offer. However, these candidates do not significantly differ from those who accepted the offer in their propensity to be recommended by the algorithm (p=0.83). Thus, the amount of unobserved selection is small and independent of performance.

6.2 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.

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 deviations lower risk tolerance on a scale from 0 to 10. Also, they report 0.5 standard

[17] We compare those recommended by the AI (we observe almost all of those in the HR treatment and all of those in the AI treatment) and those not recommended by the AI (we observe almost all of those in the HR treatment and none of those in the AI treatment). Thus, selection, shouldn't bias our results.

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 means that the effects of these inputs are not linear. For instance, age is the fourth most important variable for the algorithm, but the average age does not differ between those recommended to be hired and those who are not.

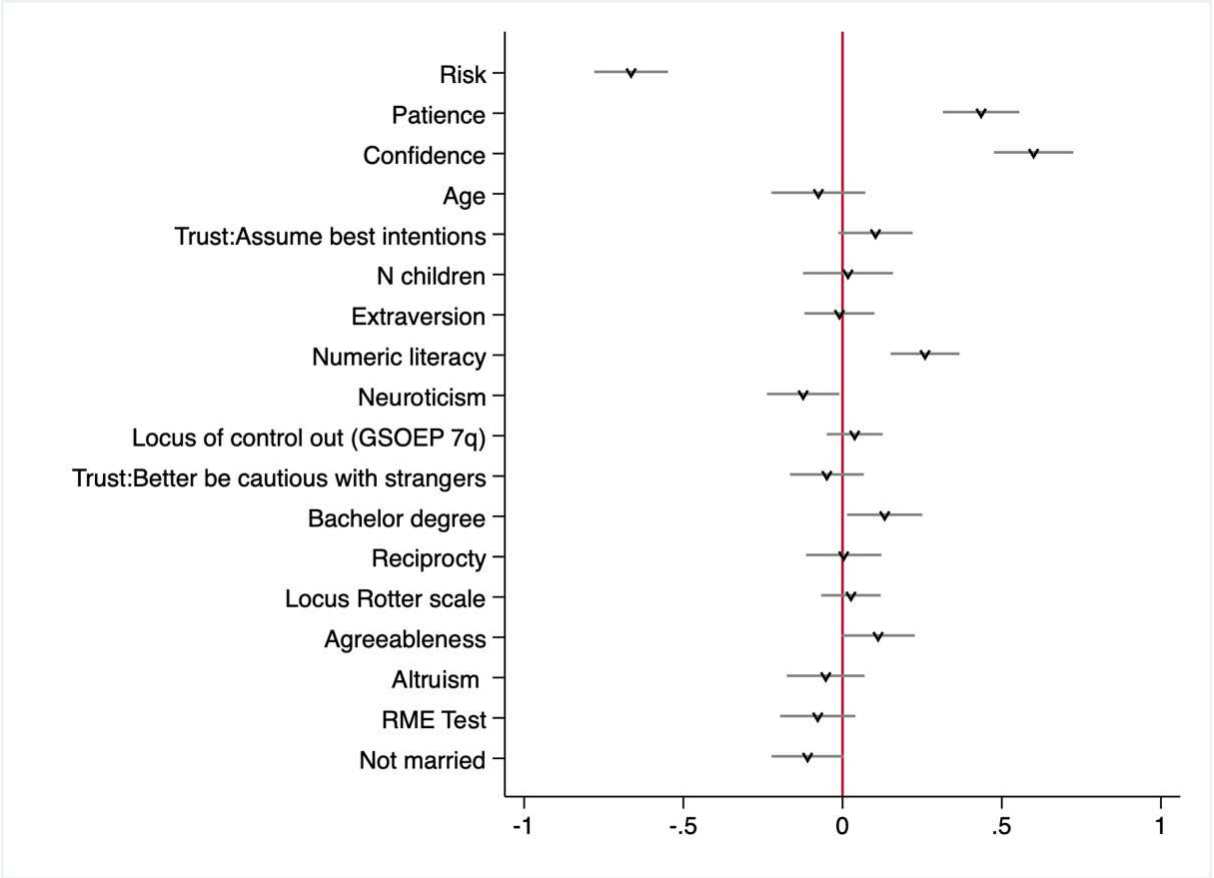


Figure 8. Coefficient of predicting recommendation to hire in OLS with the measure as outcome.
Notes: The red vertical line references zero. See notes of Table 1 for description of each measure.

6.3 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 attempt to bias the algorithm's evaluation in their favor, leading to wrong hires and thus lowering the algorithm's performance. The results of Subsection 6.1 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 are interested in 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. After pooling the answers to the survey of employees in September 2021 and of job applicants from our field experiment, 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 adding a number of 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, and not 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.

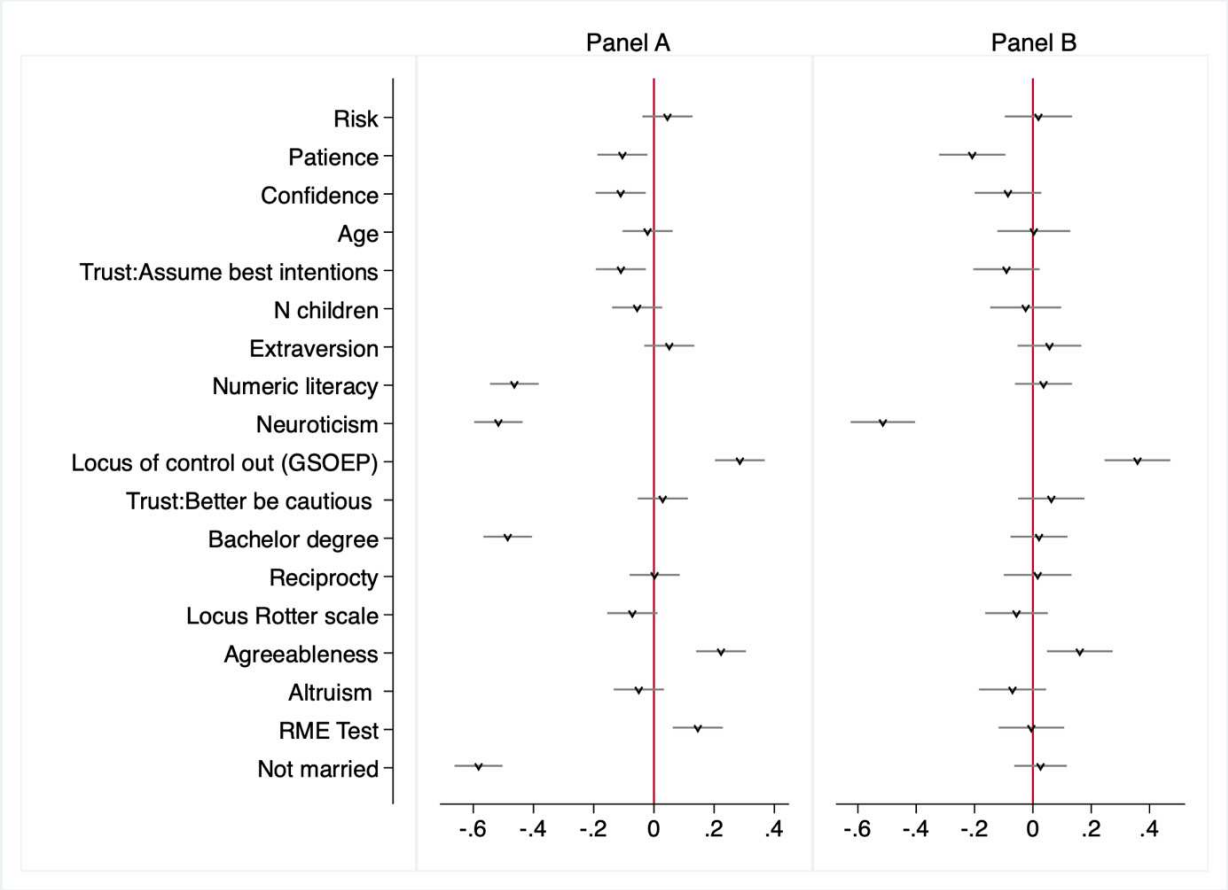


Figure 9. Coefficient plots of the OLS regression of the input variables of the algorithm on the dummy for the applicant sample relative to the employee sample.

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[18], 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

[18]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.

their knowledge. 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 are significantly different between the employee and applicant samples than without the controls. Importantly, none of the non-strategic variables significantly differ between the samples, 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 - pretend to be more patient than they are. One interpretation is that applicants perceive patience as an indicator of a low motivation or ambition. 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 in order to improve their hiring chances. For neuroticism and agreeableness, higher scores should indeed increase the chances of a hiring recommendation, as shown in the previous subsection. For locus of control, however, the overall effect is zero, since it affects the likelihood of being recommended for hiring in a non-linear 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 complement these findings while also indicating that economic measures are harder to manipulate, since their connection to productivity or employer preferences seems less predictable for candidates.

One might argue that while manipulations may initially be challenging, candidates may eventually learn to game the system. If this were the case, the proportion of candidates rejected by the algorithm should decrease over time. However, we find 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, for the duration of our experiment, we do not observe any significant pattern of candidates adapting to the algorithm.

6.4 Predicting size and riskiness of the portfolio

There are two important components of the productivity of loan officers and the likelihood to earn a bonus, namely the size of the portfolio and its quality. We study whether the algorithm recommends employees who are better sellers (have a larger portfolio) and better screeners (have a lower portfolio at risk) than others.

	PAR0, %	Portfolio size, KGS
Algorithm recommended hire	-0.011** (0.004)	274255 (680215)
Constant	0.014***	6211453***
	(0.004)	(627885)
Observations	213	213
	0.084	0.000
Clusters	92	92
Sample	Employed 02.24	Employed 02.24

Table 4. Portfolio at risk in percent and portfolio size. Standard errors are clustered at the level of offices. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4 presents the results of an OLS regression of the portfolio at risk with a delay of more than 0 days and the portfolio size. The negative coefficient of algorithmic recommendation in the first regression shows that those recommended for hiring have significantly less risky portfolios. This result aligns well with the observation that the algorithm selects employees with significantly higher self-reported risk-aversion. As can be taken from the non-significant coefficient of the dummy for the algorithmic recommendation in the second regression, there is no significant difference in portfolio size between those applicants recommended by the algorithm and others. This could mean that the algorithm does not predict selling skills well, or that self-selection takes place based solely on the volume of sales, i.e., those employees who do not sell a lot leave the company. In this case, all workers will have portfolios of comparable sizes, but the risk of the portfolio will differ, which is realized with some delay.

7 Algorithmic versus HR hiring

We now turn to the results concerning the second goal of the field experiment, namely a comparison of the effectiveness of algorithmic hiring to current HR practices of the firm. The details of the design are provided in Section 5.

7.1 Treatment effects

We first consider the impact of the treatment, i.e., being hired based on the recommendation of HR or AI, on the probability of getting a bonus. We consider a bonus payment in the 12th month of employment, our pre-registered measure, and a bonus in February 2024. Figure 10 shows that the proportion of employees who receive a bonus is higher for employees hired by the AI than for those hired by the managers, but the differences are only marginally significant (two-sided Fisher exact test p=0.07 for the probability of receiving a bonus in the 12th month of employment and p=0.055 for a bonus in February 2024, respectively).

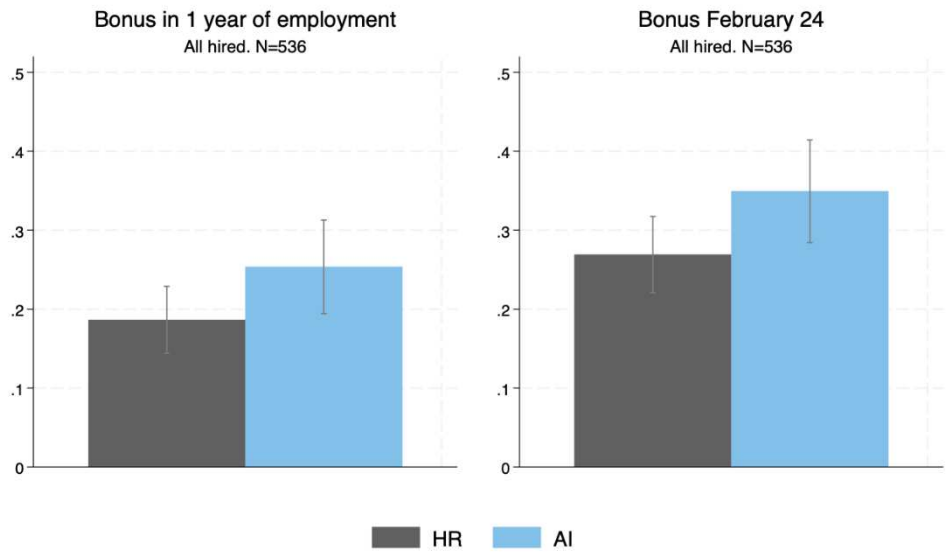


Figure 10: Treatment effect on bonus.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

When restricting the sample to workers still employed when the bonus decision was taken, the pattern is similar, as can be seen in Figure 11. The differences are statistically significant both for the probability of receiving a bonus in the 12th month of employment and for a bonus in February 2024 (two-sided Fisher exact test p=0.03 and p=0.02, respectively).

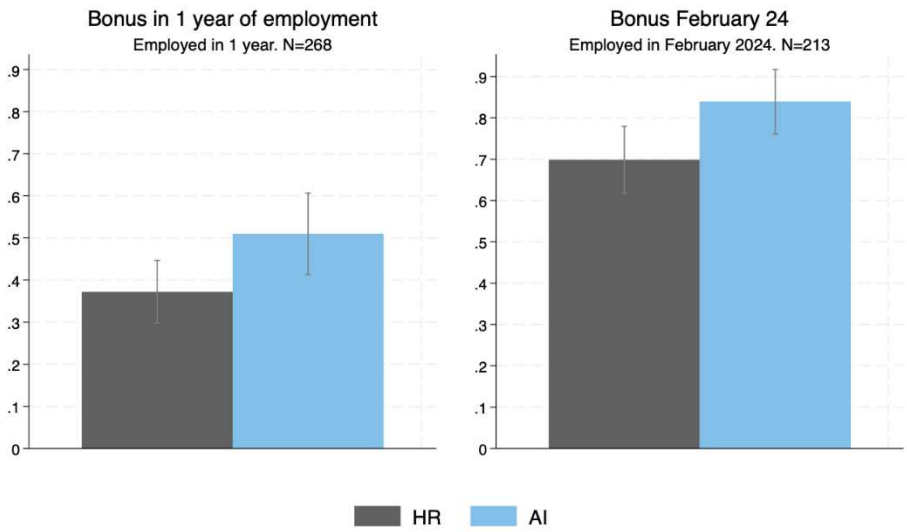


Figure 11: Treatment effect on bonus conditional on being employed.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

We run probit regressions where the dependent variables are either whether a person obtains a bonus or whether they leave the company and the independent variable indicates whether the employee was in the AI treatment rather than the HR treatment. The regression results in Table 5 indicate that candidates hired under the AI treatment are more likely to receive a bonus at both the 12-month mark and in February 2024 compared to those hired under the HR treatment irrespective of their employment status at the time of the bonus decision. However, the effect is only marginally significant (p=0.08 and p=0.07, respectively). The regression results also confirm that leaving the firm is not less likely for employees in the AI treatment than for those in the HR treatment. Finally, we find a significant positive effect of the AI treatment on the probability of receiving a bonus in the 12th month of employment and in February 2024, conditional on being employed when the bonus decision is taken.

	Bonus 1 year	Bonus 02.24	Left 1 year	Left 02.24	Bonus 1 year	Bonus 02.24
AI treatment	0.066* (0.038)	0.079* (0.043)	0.004 (0.050)	-0.036 (0.042)	0.135** (0.058)	0.144** (0.061)
Observations	536	536	536	536	268	213
Clusters	102	102	102	102	91	90
Sample	All	All	All	All	Employed	Employed

Table 5: Performance of employees. Marginal effect of probit regressions of the outcome on the AI treatment dummy, with standard errors clustered at the level of offices. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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. We conclude from our findings that algorithmic hiring is only marginally more efficient than the

current HR practice. The profitability of algorithmic hiring will depend on factors such as hiring costs, the time costs of managers and regional managers, and, importantly, the intensive margin of benefits generated by workers in the AI treatment. Unfortunately, we do not have access to this data, as it is challenging for the firm to estimate and includes confidential salary information.

7.2 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. 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. Moreover, the company organizes office-level competitions to motivate employees who have not yet reached the bonus threshold, indicating that managers value more than just bonus whether a bonus is paid or not.[19] 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 (Table B.4). 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, but the relationship is only significant when using an index indicating whether the employee received very high qualitative ratings. 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.

[19] Competitions are held every year from February to October within the company. The main incentive is to be selected for the end-of-year corporate event. Each office represents a team. Targets change each month and often focus on growth, quality or sales, but often also emphasize specific products or include metrics such as adherence to procedures or client feedback. Office rankings are announced at the end of each month.

8 Discussion and conclusions

We explored the usefulness of behavioral and psychological measures to predict the performance of employees. We trained a random forest algorithm with the data from loan officers in a microfinance firm in Kyrgyzstan and employed the algorithm to predict the performance of new employees and job applicants. A combination of psychological and behavioral measures as well as demographic variables used by predicts the performance of both new employees and applicants well, in spite of the biased training sample and some strategic responses by the applicants.

Our findings offer a positive perspective on the use of behavioral measures in hiring processes. Specifically, in 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 can learn over time how to answer the survey questions used for algorithmic hiring. However, many variables enter in a non-linear way, which makes it harder for applicants to provide optimal answers. Also, we find no evidence of a higher acceptance rate of applicants after one year that the algorithm has been used.

The experiment indicates that algorithmic hiring, when calibrated to a 30% rejection rate, only marginally outperforms local and regional managers in selecting employees. However, the cost of screening with AI may be substantially lower. In the context of our firm, hiring is just one of many responsibilities of regional and local managers. Thus, employing AI hiring tools can free up managerial resources for other tasks, such as training employees, managing client risk, and conducting anti-fraud audits. In addition, performance predictions for individual employees are needed for other managerial decisions such as promotions, where AI could also be beneficial.

The study was conducted in the specific context of a microcredit firm in Kyrgyzstan. It is uncertain how well our specific algorithm 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. 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. This is 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. The results validate and refine 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. Our approach adds the selection and weighting of these traits with the help of AI and demonstrates 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.

Ash, E., Galletta, S., & Giommoni, T. (2020). A Machine Learning Approach to Analyzing Corruption in Local Public Finances. *Center for Law & Economics Working Paper Series*, *6*.

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. Mimeo.

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

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. *arXiv preprint arXiv:2310.17688*.

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.

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., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition* 181, 21-34.

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. (2023). Strategic Responses to Personalized Pricing and Demand for Privacy: An Experiment. arXiv preprint arXiv:2304.11415.

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.

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

Dargnies, M. P., Hakimov, R., & Kübler, D. (2024). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science*.

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.

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. (2024) Strategic Information Disclosure to Recommendation Algorithms: An Experiment. Working paper

Hakimov, R., Schmacker, R., & Terrier, C. (2023). Confidence and college applications: Evidence from a randomized intervention (No. SP II 2022-209). WZB Discussion 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* (No. w31690). National Bureau of Economic Research.

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.

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.

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). Briarcliff Manor, NY 10510: Academy of Management.

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.

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.

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 (forthcoming). Hiring as Exploration. *Review of Economic Studies.*

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.

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* (No. w32889). National Bureau of Economic Research.

Appendix A

Incentivized Measures	
Risk preferences with MPL	Holt & Laury (2002)
Time preferences with MPL	Cohen et al. (2020)
Trust game with anonymous other participant	Berg et al. (1995)
Ultimatum game	Güth et al. (1982)
Altruism through donation to charity	
Non-Incentivized Measures	
General willingness to take risk, from 1 to 10	Dohmen et al. (2011)
Non-incentivized version of Gneezy and Potters risk task	Gneezy & Potters (1997)
Staircase measure of risk	Falk et al. (2018)
Self-assessed patience from 1 to 10	Falk et al. (2018)
Staircase measure of time preferences	Falk et al. (2018)
Measure of reciprocity	Falk et al. (2018)
Willingness to return a favor, from 1 to 10	Falk et al. (2018)
Willingness to punish at a cost, from 1 to 10	Falk et al. (2018)
Trust Likert scale	Falk et al. (2018)
Caution towards strangers Likert scale	Falk et al. (2018)
Altruism from 1 to 10	Falk et al. (2018)
Big 5 personality traits	GSOEP,Goldberg (1992)
4-item Rotter Internal-External Locus of Control Scale	McGee et al. (2016)
7-item Internal Locus of Controls,	GSOEP
10-item Grit	Duckworth et al. (2007)
10-item Reading the Mind in the Eyes Test (RMET)	Weidmann et al. (2021)
Wonderlic test	Dodrill (1981)
6-question numeric literacy test, ELSA	
CRT	Frederick (2005)
Local network measure	Number of cousins living nearby
Confidence in answers to ELSA+CRT	
Relative confidence compared to colleagues	

Survey

Welcome to the questionnaire. This is a part of a research study, so by being attentive and answering honestly, you will help science and management. In some of the questions, you will be able to earn money, and one of these questions will be randomly chosen, and your earnings will be paid to you on your bank card. Some other questions do not have monetary payoffs but allow us to know you better, so answer honestly. None of your colleagues or managers will find out your answers, but based on them, we might be able to provide individual advice to you, so it is best for you to answer honestly.

Incentivized

Risk preference

- Would you rather receive a certain payment of 100€ or participate in a lottery with 50% chance of having 200€ and 50% chance of having 0€ ?
- … (change amounts and probabilities for other questions, also see Ordered Selection System with the circles in Jagelka)

Time preference

- Would you rather receive 100€ today or 120€ in 12 months (… etc)

Trust

You are in a situation where you are given 10€. Now, you have to decide to send an amount between 0 and 10 to an anonymous second player. The amount the second player receives will be tripled by the experimenter.

After you made your choice, the second player will also have to choose an amount between 0 and 10 to send back to you. What amount do you choose?

Positive reciprocity

- You are in the opposite situation as before. You first receive an amount sent by the first player that has been tripled by the experimenter. You now have to choose an amount between 0 and 10 to send back to the first player that will also be tripled. What amount do you choose?

Negative reciprocity

Imagine the following situation: together with a person whom you do not know, you won 100 Euro in a lottery. The rules stipulate the following: One of you has to make a proposal about how to divide the 100 Euro between you two. The other one gets to

know the proposal and has to decide between two options. He or she can accept the proposal or reject it. If he or she accepts the proposal, the money is divided according to the proposal. If he or she rejects the proposal, both receive nothing. Suppose that the other person offered the following splits:

50 Euro for you and 50 Euro for himself/herself. Do you accept this split?

- If you do, you will receive 50 Euro and the other person will receive 50 Euro. If you reject, both of you receive 0 Euro.

40 Euro for you and 60 Euro for himself/herself. Do you accept this split?

30 Euro for you and 70 Euro for himself/herself. Do you accept this split?

20 Euro for you and 80 Euro for himself/herself. Do you accept this split?

10 Euro for you and 90 Euro for himself/herself. Do you accept this split?

Altruism

- If you are endowed with €100, how much of this endowment would you give to a charitable organization?

Non-incentivized

Risk preference

Use a scale from 0 to 10, where 0 means "completely unwilling to do so" and 10 means "completely willing to do so":

- How willing or unwilling you are to take risks?[20]

Please consider what you would do in the following situation: Imagine that you have won 100,000 Euros in a lottery. Almost immediately after you collect the winnings, you receive the following financial offer from a reputable bank, the conditions of which are as follows: There is the chance to double the money within two years. It is equally possible that you could lose half of the amount invested. You have the opportunity to invest the full amount, part of the amount or reject the offer. What share of your lottery winnings would you be prepared to invest in this financially risky, yet lucrative investment?

- 100 000
- 80 000
- 60 000
- 40 000

[20] Falk (2016). The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences. http://ftp.iza.org/dp9674.pdf

- 20 000
- Nothing.

Staircase measure (Falk et al 2016)

Please imagine the following situation: You can choose between a sure payment and a lottery. The lottery gives you a 50 percent chance of receiving 300 Euro. With an equally high chance you receive nothing. Now imagine you had to choose between the lottery and a sure payment. We will present to you different situations. The lottery is the same in all situations. The sure payment is different in every situation.

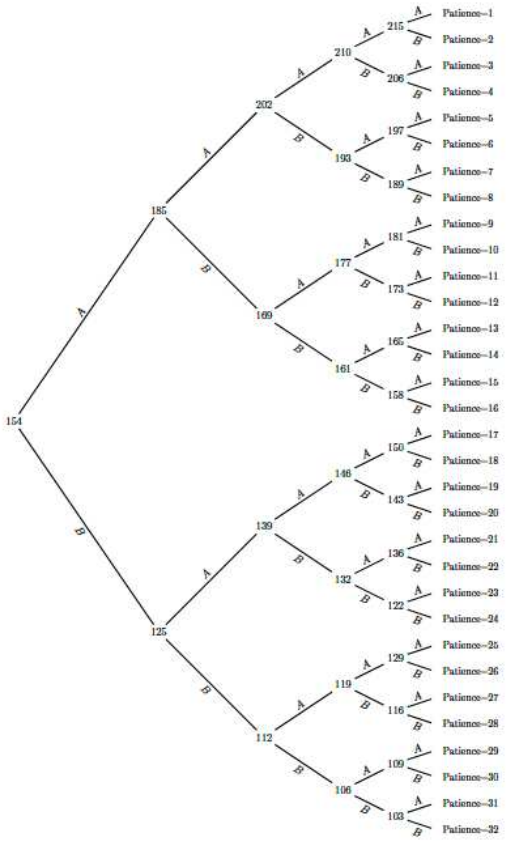


Figure 3: Tree for the staircase time task (numbers = payment in 12 months, A = choice of "100 euros today", B = choice of "x euros in 12 months"

Time preference

Use a scale from 0 to 10, where 0 means "does not describe me at all" and 10 means "describes me completely"

- Would you describe yourself as a patient person?

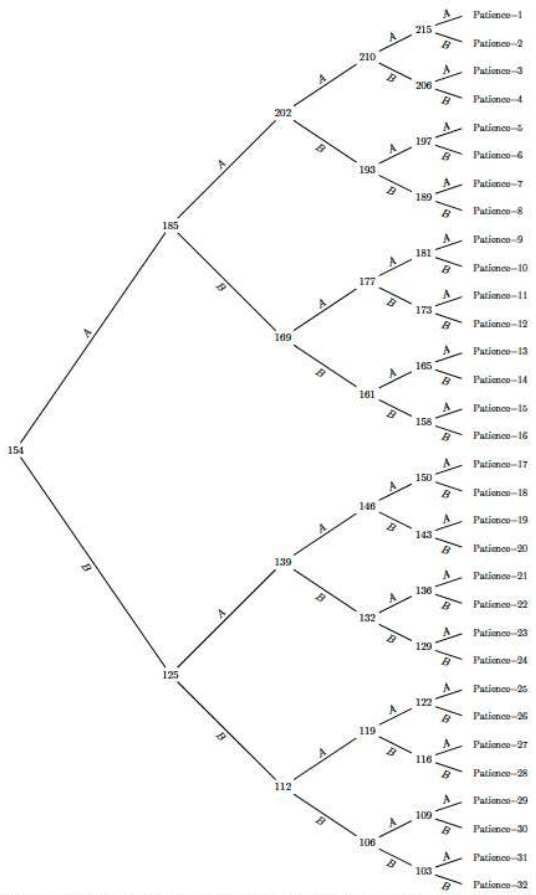
Use a scale from 0 to 10, where 0 means “completely unwilling to do so” and 10 means “ completely willing to do so” (Becker):

- How willing are you to give up something that is beneficial for you today in order to benefit more from it in the future?

Staircase measure (Falk et al 2016)

Suppose you were given the choice between receiving a payment today or a payment in 12 months. We will now present to you five situations. The payment today is the same in each of these situations. The payment in 12 months is different in every situation. For each of these situations we would like to know which one you would choose. Please assume there is no inflation, i.e., future prices are the same as today’s prices. Please consider the following: Would you rather receive amount 100 today or x in 12 months?

Figure A1: Hypothetical Choices: Staircase Method



Notes: Tree for the Immediate-Delay staircase task (numbers = payment in 12 months). A = choice of “100 GEL today”, B = choice of “x euros in 12 months”. The staircase procedure worked as follows. First, each respondent was asked whether they would prefer to receive 100 GEL today or 154 GEL in 12 months from now (leftmost decision node). In case the respondent opted for the payment today (“A”), in the second question the payment in 12 months was adjusted upwards to 185 GEL. If, on the other hand, the respondent chose the payment in 12 months, the corresponding payment was adjusted downward to 125 GEL. The last column indicates the coding of patience based on the participant’s decisions. The tree for Delay-Delay follows the same procedure with A = choice of “100 GEL in 12 months”, B = choice of “x euros in 24 months”.

Reciprocity

Imagine the following situation: you are in an unfamiliar city and realize you lost your way. You ask a stranger for directions. The stranger offers to take you with their car to your destination. The ride takes about 20 minutes and costs the stranger about 20 Euro in total. The stranger does not want money for it. You carry six presents with you. The cheapest present costs 5 Euro, the most expensive one 30 Euro.

Do you give one of the presents to the stranger as a "thank-you"-gift? If so, which present do you give to the stranger? You can choose from the following options: Give nothing or the present of 5, 10, 15, 20, 25, or 30 Euro) (Falk et al. 2016)

Use a scale from 0 to 10, where 0 means “completely unwilling to do so” and 10 means “completely willing to do so” (Becker):

- How willing are you to return a favour if someone did you one?
- How willing are you to punish someone who treats you unfairly, even if there may be costs for you?
- How willing are you to punish someone who treats others unfairly, even if there may be costs for you?

Trust

Answer these statements using a scale from 0 to 10, where 0 means “I completely disagree” and 10 means “I completely agree”:

- As long as I am not convinced otherwise, I assume that people only have the best intentions. (Falk et al 2016)
- When dealing with strangers it is better to be cautious. (Becker)

Altruism

Use a scale from 0 to 10, where 0 means “completely unwilling to do so” and 10 means “ completely willing to do so”

- How willing are you to give to good causes without expecting anything in return? (Falk et al 2016)
-

Big Five BFI-S 15 items, as in GSOEP

Please choose the ranking for each of the following questions. The rank should be between 1 =“does not apply to me at all” to 7 = “applies to me perfectly”.

I see myself as someone who …

- *does a thorough job.*
- *is communicative, talkative.*
- *is sometimes somewhat rude to others.*
- *is original, comes up with new ideas.*
- *worries a lot.*
- *has a forgiving nature.*
- *tends to be lazy.*
- *is outgoing, sociable.*
- *values artistic experiences.*
- *gets nervous easily.*
- *does things effectively and efficiently.*
- *is reserved.*
- *is considerate and kind to others.*
- *has an active imagination.*
- *is relaxed, handles stress well.*

Locus-of-control (Rotter)

Abbreviated 4-item Rotter Internal-External Locus of Control Scale (McGee&McGee 2016)

A. What happens to me is my own doing.

B. Sometimes I feel that I don't have enough control over the direction my life is taking.

A. When I make plans, I am almost certain that I can make them work.

B. It is not always wise to plan too far ahead because many things turn out to be a matter of good or bad fortune.

A. In my case getting what I want has little or nothing to do with luck.

B. Many times we might just as well decide what to do by flipping a coin.

A. Many times I feel that I have little influence over the things that happen to me.

B. It is impossible for me to believe that chance or luck plays an important role in my life.

Locus of control. SOEP 7 item

Please choose the ranking for each of the following questions. The rank should be between 1 ="does not apply to me at all" to 7 = "applies to me perfectly"

How my life goes depends on me (Internal LoC)

If a person is socially or politically active, he/she can have an effect on social conditions (Internal LoC)

One has to work hard in order to succeed (Internal LoC)

Compared to other people, I have not achieved what I deserved (External LoC)

I frequently have the experience that other people have a controlling influence over my life (External LoC)

The opportunities that I have in life are determined by the social conditions (External LoC)

I have little control over the things that happen in my life (External LoC)

Grit (Duckworth)

Here are a number of statements that may or may not apply to you. There are no right or wrong answers, so just answer honestly, considering how you compare to most people. Answer these statements using this scale:

- Very much like me
- Mostly like me
- Somewhat like me
- Not much like me
- Not like me at all

1. New ideas and projects sometimes distract me from previous ones.
2. Setbacks don't discourage me. I don't give up easily.
3. I often set a goal but later choose to pursue a different one.
4. I am a hard worker.
5. I have difficulty maintaining my focus on projects that take more than a few months to complete.
6. I finish whatever I begin.
7. My interests change from year to year.
8. I am diligent. I never give up.
9. I have been obsessed with a certain idea or project for a short time but later lost interest.

10. I have overcome setbacks to conquer an important challenge.

Reading the Mind in the Eyes Test (RME)

Question 8

Despondent

Relieved

Shy

Excited

Question 9

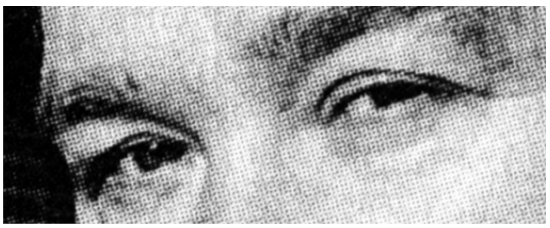
Annoyed

Hostile

Horrified

Preoccupied

Question 12



Indifferent

Embarrassed

Sceptical

Dispirited

Question 14

Irritated

Disappointed

Depressed

Accusing

Question 15

Contemplative

Flustered

Encouraging

Amused

Question 19

Arrogant

Grateful

Sarcastic

Tentative

Question 22

Preoccupied

Grateful

Insisting

Imploring

Question 24

Pensive

Irritated

Excited

Hostile

Question 32

Serious

Ashamed

Bewildered

Alarmed

Question 36

Ashamed

Nervous

Suspicious

Indecisive

Cognitive ability as in ELSA (Numeracy test) and CRT (Bortolotti et al. 2020)

In the following block, you will answer 9 questions, please answer as many of them as you can within 3 minutes. (should be on one screen, time counted).

The Numeracytest administered in the individual survey included the following six questions

1. If you buy a drink for 85 cents and pay with a one-euro coin, how much change should you get?
2. In a sale, a shop is selling all items at half price. Before the sale a sofa costs 300 euros. How much will it cost in the sale?
3. If the chance of getting a disease is 10 per cent, how many people out of 1,000 would be expect to get the disease?
4. A second-hand car dealer is selling a car for 6,000 euros. This is two-thirds of what it cost new. How much did the car cost new?
5. If 5 people all have the winning numbers in the lottery and the prize is 2 million, how much will each of them get?
6. Let's say you have 200 in a savings account. The account earns ten per cent interest per year. How much will you have in the account at the end of two years?

The Cognitive Reflection Test administered in the individual survey included three questions adapted from Frederick (2005)

1. A bat and a ball cost 1.10 euros in total. The bat costs 1.00 euros more than the ball. How much does the ball cost?
2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

Measure of local network

How many siblings and cousins, counted together live no further than 15 km from your house?

Confidence

How many out of the nine questions you think you answered correctly?

Imagine we rank performances of you and 99 random colleagues who answered this test in the nine questions from the highest number of correct answers (place 1) to the lowest number of correct answers (place 100), which place do you think you have?

Appendix B. Additional algorithms comparisons and results

Table B1: Confusion matrix for the prediction task based on 250 runs of the algorithm using firm data only

	Bonus	No bonus
Predict bonus	56%	32%
Predict no bonus	3%	9%

Table B2: Confusion matrix for the prediction task based on 250 runs of the algorithm using firm data and non-incentivized survey measures

	Bonus	No bonus
Predict bonus	61%	27%
Predict no bonus	3%	9%

Figure B1 presents the density of the predicted probabilities conditional on having received a bonus or not. To generate these graphs, we ran each model 250 times. Each time, individuals in the validation group were assigned a recorded probability of receiving a bonus. Since all 674 employees were part of the validation sample at some point, we calculated the average probability for each employee based on all instances they were included in the validation group. The density plots are presented and grouped by the actual status of whether they received a bonus. The densities are more separated in the right panel, confirming the better performance of the model that includes non-incentivized measures.

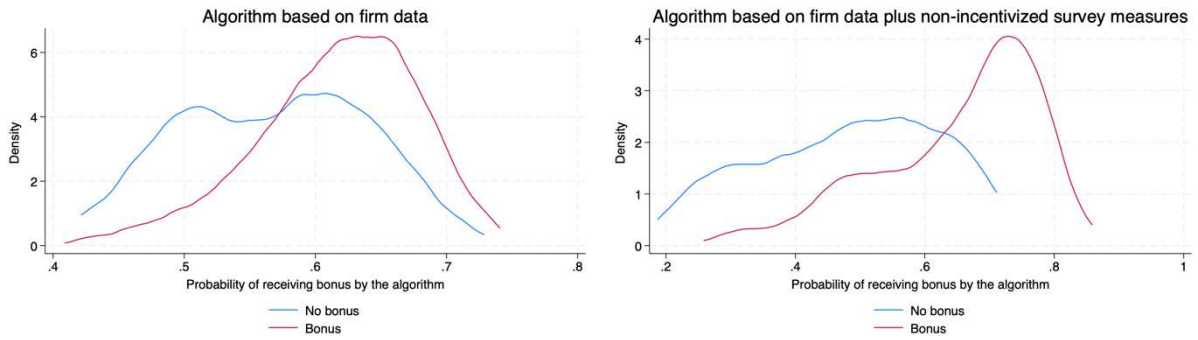


Figure B1. The density of the predicted probabilities conditional on having received a bonus or not. Left panel: The algorithm based on firm data alone. Right panel. The algorithm based on firm and non-incentivized data.

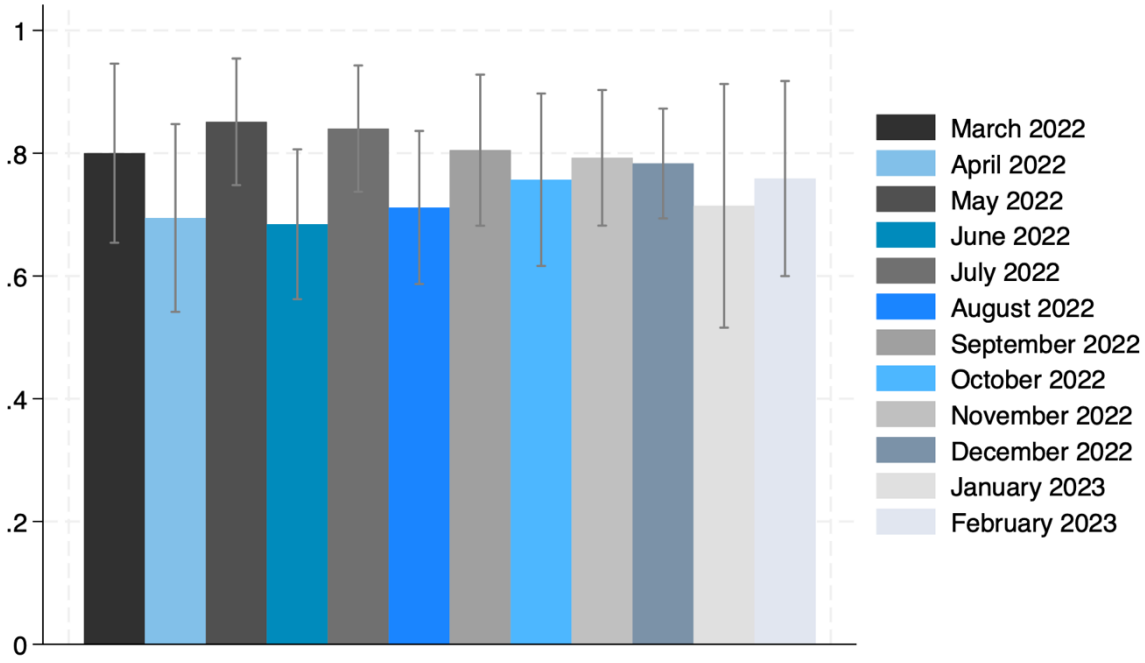


Figure B2. Proportion of candidates recommended for hiring by algorithm.

Robustness: Continuous measure of algorithmic recommendation

The performance of the algorithm's recommendation depends on the selected threshold for the recommendation. Note that we aimed for a 30% rejection rate, but this is an arbitrary threshold based on the top management's recommendation. However, we also observe a more continuous measure for each applicant in the sample. We run the algorithm 250 times, varying the split between the training sample and the validation set in each iteration. As a result, the algorithm classified each applicant as either productive (predicted to receive a bonus) or non-productive in every run. This process generated a score for each applicant, ranging from zero to 250. As a robustness check, instead of using the binary variable indicating whether the algorithm recommended to hire an applicant, we can use the score, i.e., the proportion of runs in which the algorithm recommended hiring. Table B.3 presents results analogous to Table 3 using the continuous variable of the percentage of hiring recommendations by the algorithm instead of the binary measure.

	Bonus 1 year	Bonus 02.24	Left 1 year	Left 02.24	Bonus 1 year	Bonus 02.24
	0.206***	0.198***	-0.082	-0.113**	0.331***	0.271***

% of runs where algorithm recommends hiring	(0.043)	(0.045)	(0.053)	(0.049)	(0.067)	(0.060)
Observations	536	536	536	536	268	213
Clusters	95	102	95	102	91	90
Sample	All	All	All	All	Employed	Employed

Table B.3: Performance of employees. Marginal effects of probit regression of the outcome on the dummy for the algorithm predicting a bonus, with standard errors clustered at the level of offices. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

The marginal effects of the probit regressions reported in Table B.3 are generally stronger and more precisely estimated than in Table 3. Thus, the proportion of recommendations contains important cardinal information about the quality of the applicants, implying that the cutoff decision for hiring could be optimized.

Do local managers and the algorithm value the same skills?

We have compared algorithmic and HR hiring in selecting high-productivity candidates. However, productivity alone does not necessarily define a good employee. The algorithm may prioritize performance over collegiality and teamwork or select candidates whom managers find unsuitable for other reasons. We now attempt to assess how algorithmically selected employees compare to others in 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. All managers were asked to answer four questions about each loan officer in their office:

1. On a scale from 1 to 10, how would you rate the employee, from your personal perspective?
2. On a scale from 1 to 10, how would you rate the employee's contribution to the office's success? For instance, think of the between-office competitions in 2023.
3. On a scale from 1 to 10, how would you rate the employee's contribution to creating a collegial atmosphere in the office?
4. How likely (0 to 100%) do you think the employee is to still be working at the firm in one year from now (0 for sure not, 100 for sure yes)?

In addition to the raw answers to each question, we generated an index from the four questions, which ranges from 0 to 4. For each question, the index increases by 1 if the employee received a 9 or 10 in questions 1 to 3 and 95 or above in question 4. Due to the skewed distribution of

answers in favor of extreme values, the index shows in how many of the 4 questions the employee received an almost maximum score. Of the employees evaluated by managers in January 2024, 208 participated in our hiring experiment, and all were still employed at the firm as of February 2024.

Table B.4 presents the results of an OLS regression where the dependent variables are the scores of the employees for each question as well as the index. We analyze whether the scores are correlated with the propensity of the employees to receive a bonus and with the algorithmic recommendation.

	Q1 Personal rating	Q2 Office success contribution	Q3 Positive atmosphere	Q4 Probability of staying	Index of maximum scores
Bonus in 02.24	0.680*	0.396	0.200	6.112	0.722**
	(0.373)	(0.415)	(0.319)	(7.757)	(0.305)
Algorithm recommended hire	-0.124	-0.116	0.121	-4.102	0.098
	(0.348)	(0.383)	(0.398)	(7.012)	(0.326)
Constant	7.831***	7.438***	8.436***	70.430***	1.450***
	(0.329)	(0.383)	(0.335)	(7.374)	(0.315)
Observations	208	208	208	208	208
R^2	0.025	0.007	0.004	0.005	0.048
Sample	Employed 02.24	Employed 02.24	Employed 02.24	Employed 02.24	Employed 02.24

Table B.4. Employee scores from local manager.
Notes: *Bonus in 02.24 is a dummy variable indicating whether an individual received a bonus in February 2024. Algorithm recommended hire is a dummy variable representing the original recommendation made by the algorithm at the hiring stage. The standard errors clustered on the level of each office.* * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results in Table B.4 show that the scores of employees provided by local managers are consistently higher for employees who receive a bonus than for those who do not, but the relationship is only significant for the index and marginally significant for the personal rating by the local manager. In contrast, the scores are not significantly associated with the recommendation of the algorithm.[21] Thus, despite the fact that the algorithm is trained to predict a bonus, it does not select applicants who fare significantly worse on team work and social skills.

[21] Note that controlling for the bonus in February 2024 does not drive the result. In the model without this control, the algorithm recommendation also does not significantly correlate with any of the outcomes. Also, models with AI treatment instead of the algorithm recommendation lead to not significant treatment effects.