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

Marie-Pierre Dargnies (University of Lille)

Rustamdjan Hakimov (University of Lausanne)

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

February 2026

Abstract

The adoption of Artificial Intelligence (AI) for hiring processes is often impeded by a scarcity of comprehensive employee data. We hypothesize that the inclusion of behavioral measures elicited from applicants can enhance the predictive accuracy of AI in hiring. We study this hypothesis in the context of microfinance loan officers, running an RCT where AI makes hiring decisions for half of the applicants. Our findings suggest that survey-based behavioral measures markedly improve the predictions of a random-forest algorithm trained to predict productivity within sample relative to demographic information alone. We then validate the algorithm's robustness to the selectivity of the training sample and potential strategic responses by applicants by running two out-of-sample tests: one forecasting the future performance of novice employees, and another with a field experiment on hiring. Both tests corroborate the effectiveness of incorporating behavioral data to predict performance and the usefulness of AI hiring.

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

[1]We thank Brice Corgnet, Guido Friebel, Johannes Johnen, Amma Panin, Christian Zehnder, and participants of the 2025 European Decision Sciences Research Day at INSEAD (Fontainebleau), the 2023 Zurich Workshop on Economics and Psychology, the CREST workshop on experimental economics 2023, the Heilbronn Workshop on Field Experiments in Economics and Business 2024, and the Newcastle Experimental Economics Workshop 2024, seminar participants at University of Lille, Bamberg, Lingnan University Hong Kong, Lund University, NUS Singapore, 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 machine learning in particular, has progressed rapidly in recent years (Goodfellow et al. 2016, Bengio et al. 2023). While earlier automation primarily affected jobs with a high share of routine tasks (Autor 2015, Arntz et al. 2016), machine-learning systems are increasingly applied to screening and prediction problems that traditionally relied on human judgment. Across several domains, algorithmic predictions can match or outperform expert decision-making, including in contexts related to selection and personnel decisions (e.g., McKinney et al. 2020, Mullainathan and Obermeyer 2022, Kleinberg et al. 2018, Ash et al. 2020, Chalfin et al. 2016). In labor markets, the diffusion of electronic applications has expanded applicant pools and increased the demand for scalable screening, contributing to the rapid spread of vendor-provided tools that score CVs and asynchronous interviews.[2] A central concern is fit: these largely 'one-size-fits-all' tools are typically developed and trained outside the client firm, so their effectiveness and fairness need not transfer across firms, jobs, and local labor markets. Consistent with this concern, audits document limited external stability—small changes in the format of input to measure applicants' personality (e.g., CV versus LinkedIn) lead to changes in AI output (Rhea et al. 2022 and references therein).

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

[2] A vast majority of firms is already using some kind of AI tools for screening applicants, see https://www.forbes.com/sites/janehanson/2023/09/30/ai-is-replacing-humans-in-the-interview-processwhat-you-need-to-know-to-crush-your-next-video-interview/ (last accessed on 9.11.2025).
The market for such tools is predicted to expand further, https://www.maximizemarketresearch.com/market-report/global-ai-recruitment-market/63261/?utm_source=chatgpt.com(last accessed on 20.02.2024), with new methods emerging continuously (e.g., Guenzel et al. 2026).

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

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

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

In an out-of-sample test, a random forest algorithm, trained only on firm data such as age, education level, marital status, and other characteristics, correctly classifies about 65% of employees. Adding the non-incentivized measures significantly improved the performance of the

[3]In addition, even accurate tools may face implementation constraints because managers, clients, and workers can be averse to algorithmic decision-making or override recommendations (Highhouse 2008, Dietvorst et al. 2015, Dargnies et al. 2024).

random forest algorithm by five percentage points. A similar improvement was observed with the use of incentivized measures. However, the inclusion of incentivized measures alongside non-incentivized ones did not further enhance the algorithm's performance, suggesting that the two sets of measures are substitutes. Because non-incentivized measures are easier to elicit in practice, the remaining analyses and the field experiment use the algorithm trained on firm data and the non-incentivized measures.

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

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

by local and regional managers was overridden if it conflicted with the algorithm's recommendation.

To address the first goal of the experiment—evaluating the effectiveness of algorithmic hiring relative to the firm's HR practices—we compared employee performance under the two experimental treatments at the 12-month mark. During the experiment, managers rejected far fewer candidates than anticipated, which reduces treatment variation and limits statistical power of the experiment.[4] Thus, we are comparing a sample with screening of applicants by AI with an almost non-screened sample. Employees hired in the AI treatment were more likely to receive a bonus and exhibited higher productivity than those hired in the HR treatment, although these differences become only marginally significant once we correct for multiple hypotheses. Beyond performance indicators such as bonuses and productivity, a key outcome for the firm is the profitability of newly hired employees. Profitability shows a clearer pattern than productivity: average profitability at 12 months is significantly higher in the AI treatment, and despite fewer hires overall, the total profit generated by AI-hired employees over the year by far exceeds that of the HR treatment. Taken together, these findings indicate that, although differences on pre-registered outcomes are only marginally significant, the algorithm outperformed the firm's traditional hiring process.

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

In the AI treatment, hiring decisions follow the algorithm's recommendation, allowing us to observe job performance for a representative subset of candidates positively evaluated by the algorithm. By contrast, in the HR treatment, managers' near-universal acceptance implies that many candidates who would have been rejected by the algorithm were nonetheless hired. This

[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.

feature of the data gives us access to a quasi-representative sample of algorithm-rejected candidates whose subsequent job performance can be observed. Together, these two treatments allow us to evaluate the algorithm's predictive power on both sides of the decision threshold, substantially mitigating concerns related to selective labels.[5] Applicants recommended by the algorithm were significantly more likely to receive a bonus, had higher productivity, and held portfolios with fewer delayed loans at the 12-month mark relative to those not recommended. In contrast, we find no robust evidence of differences in retention at 12 months. Overall, these results indicate that the algorithm is robust to selection effects and to potential strategic manipulation of survey responses, although its predictive power for portfolio size, number of issued loans, and retention is weaker in the applicant sample than in the employee sample. We investigate 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.[6] At the same time, the proportion of applicants recommended for hiring by the algorithm does not increase over time in the one-year period of the experiment. If (successful) gaming occurs, this proportion should rise. Thus, while some signs of manipulation attempts exist, participants failed to learn how to manipulate the algorithm within the period of one year that the experiment lasted.

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

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

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

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

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

Related literature

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 psychological as well as economic measures (all non-incentivized) can complement each other. We also provide a novel method to investigate and compare the measures' robustness to strategic manipulation, yielding 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, which includes contributions on the impact of technology and other procedures on hiring practices. Awuah et al. (2025) experimentally compare human screening, AI-augmented human screening, and fully automated pre-interview evaluation using ChatGPT-4.0 in the context of teacher hiring. ChatGPT-4.0 outperforms both humans and augmented humans in predicting subsequent performance—evidence that aligns with our finding that AI can outperform discretionary screening. Autor and Scarborough (2008) show that job testing improves selection without reducing 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) train their AI in a data-rich environment, namely public sector hiring and tenure decisions and demonstrate potential of AI to outperform human decisions. We go one step further by running an RCT on algorithmic hiring.

Four studies consider the effect of AI hiring on labor demand and supply. In a field experiment, Avery et al. (2023) 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) conduct online experiments with human participants acting as job applicants examining how AI adoption and the debiasing of both human decision-makers and algorithms influence 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. Jabarian and Henkel (2025) show that interviews conducted by AI rather than humans, with candidate scoring remaining human-based, lead to more job offers and job starts, driven by the more structured content of the interviews. 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. (2023) and Corgnet (2023). The first two articles show that people have reservations against AI hiring while the last two discover more negative reactions to humans 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. Bó 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 (2025) 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, using not only psychological but also 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 the relationship 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 find a significant correlation between the two. Vieider et al. (2013) replicate this result in a large sample of 30 countries. Lönnqvist et al. (2015) examine the intertemporal stability of a survey-based measure of risk attitudes and compare it with an incentivized task developed by Holt et al. (2002). They find that the survey question performs better. Falk et al. (2018) explore non-incentivized measures for various economic behaviors, such as time preference, trust, and others. Hackethal et al. (2023) conducted a series of experiments comparing incentivized and non-incentivized risk elicitation methods and conclude that incentives do not significantly affect the results. On the other hand, Chapman et al. (2025) criticize the validation of qualitative measures with experiments theoretically and

empirically. They find that qualitative self-assessments are often correlated with multiple incentivized measures, indicating that it is not clear what they measure. 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. These findings do not rely on interpretations of what the qualitative measures exactly capture, but only take them as proxies for productivity.

2 Research design

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

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

Whenever an officer's productivity exceeds their fixed salary, the difference is paid as a bonus. The incentive intensity of this system is high: on average in 2023, the total amount of bonuses paid to loan officers was 2.7 times higher than total salaries, and top performers could receive bonuses up to the equivalent of 25 monthly base salaries of around $200. This strong pay–performance link generates substantial heterogeneity in earnings and career trajectories, with productivity strongly increasing in tenure due to returning clients who obtain larger loans and are lower risk. Conceptually, the firm's compensation scheme reflects a standard principal–agent framework in which performance-based pay aligns employee incentives with the firm's objectives.

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

The study consists of three stages

Stage 1. Collecting behavioral measures and training the algorithm

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

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

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

We anonymously match the survey responses to the personnel data of the firm that include measures of productivity of the employees. These measures are the portfolio, the portfolio at risk, the portfolio without delayed payments, the number of new loans issued and whether the employee qualified for a bonus in addition to their fixed salary. The latter is directly related to an internal measure of individual productivity. The primary goal of the management when selecting new employees is to identify those who will qualify for a bonus, and the number of salespeople in a local manager's office who have obtained a bonus is a 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. Of the 1042 employees, 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. 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 reception of a bonus within one year of employment. 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 firm data (such as age and education), another algorithm uses both firm data and the answers to the non-incentivized questions, and a third algorithm uses firm data, the answers to the non-incentivized questions, and to the incentivized questions. Since non-incentivized and incentivized measures turn out to improve the algorithm and are substitutes, we choose the algorithm which uses firm data and the easy-to-elicit non-incentivized measures in the following stages.

Stage 2. Predicting the performance of employees with short tenure

The algorithm trained on firm data and the non-incentivized measures is employed to predict the performance of the employees who have been with the firm for less than one year when they answered the questionnaire. We assess their performance one year after the survey and evaluate the 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 field experiment has three goals. First, it enables us to evaluate the firm's current HR practices against algorithmic hiring. Second, we can test the robustness of the algorithm to sample selection. Third, the experiment allows us to investigate whether potential strategic responses from candidates undermine the efficiency of the algorithm. For the experiment, some employees were hired following the normal procedure of the firm (applicants are interviewed by the local office manager and by a senior manager, i.e., the manager of a region) while others were hired following the recommendation of the AI algorithm.[7] The cutoff of the algorithm for hiring an applicant or not was determined based on the selectivity of the current HR procedure.

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

Normally, the hiring is done by the local managers (managers of the office), with the approval of senior management (managers of the region). Local managers were informed that there is a new step in the application--the survey--and that they will not be informed about the answers of the applicants. Managers continued to make their decisions based on interviews, without access to the survey responses. Thus, our experiment does not study whether the manager or the algorithm makes better use of identical information, but rather who makes better hiring decisions based on different sets of information. For the randomly determined half of the sample in the AI treatment, the recommendation was 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.[8] For the other half of the applicants, i.e., those in the HR

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

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

treatment, senior management followed their usual decision process without receiving the recommendation of the algorithm. Note that due to this design, for every applicant we know the hiring recommendation by HR *and* by the algorithm.

Neither the applicants nor the local managers were aware of participating in a study. Everyone knew about the survey in the first stage, but they were not informed of its role for the algorithm's recommendations or that this recommendation was followed for some applicants but not for others. Only the board knew about the experiment. Applicants were notified of the outcome of their application—whether they were offered a position or not. The data from the survey were only available to the researchers.

The study was approved by the IRB of LABEX, University of Lausanne, and the legal team of the firm. The study was pre-registered at the AEA RCT Registry (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 an information gain (entropy) criterion for node splitting.

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

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

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

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

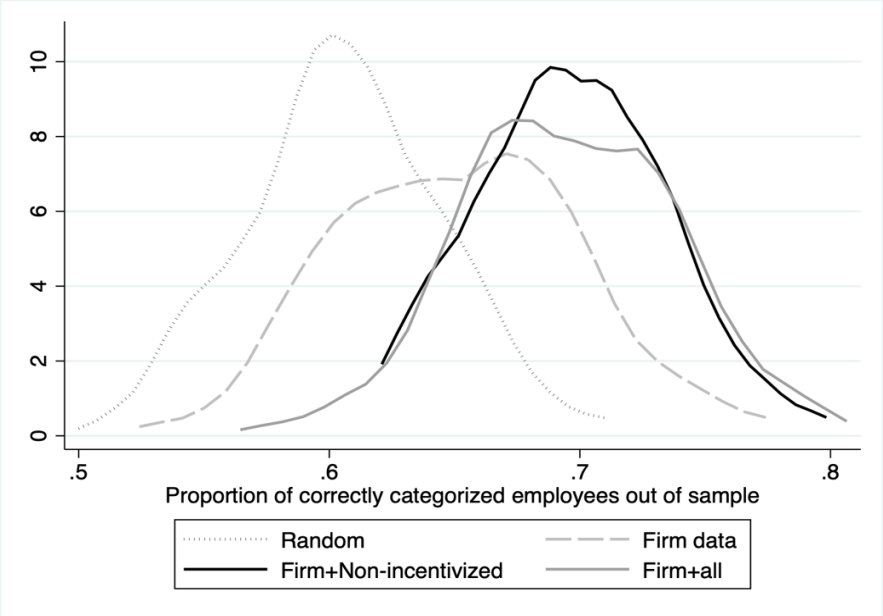


Figure 1: Validation accuracy of random-forest algorithms depending on set of variables used.
Notes: Accuracies are computed on the validation sample across 250 repeated random train/validation splits (550/124 and represent proportion of employees correctly categorized in the validation set of 124. The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

Figure 1 presents the distribution of validation accuracy of random forest algorithms across the 250 repeated splits for models estimated on different sets of explanatory variables. Using firm data only, the random forest correctly classifies 65.1% of employees in the 124-person validation sample on average. This is significantly better than an algorithm based on ten randomly generated variables (a two-sided Fisher's exact test yields $p<0.001$). Adding the selected non-incentivized

survey measures increases average validation accuracy to 69.7% ($p<0.001$). However, adding the incentivized survey measures to the algorithm that uses both the firm data and the non-incentivized survey measures does not further improve accuracy ($p=0.48$).[9] The firms' use of incentivized measures may not be feasible for logistical and financial reasons, and the results show that this is not an obstacle to obtaining an algorithm with good predictive power.

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

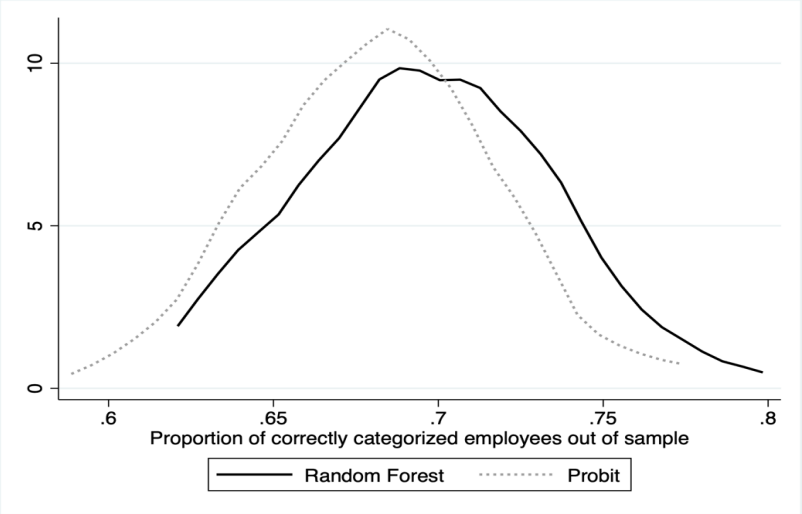


Figure 2: Validation accuracy of the random forest and probit regression. *Notes: Accuracies are computed on the validation sample across the repeated splitting protocol described above. The feature set corresponds to the final set of selected non-incentivized measures.*

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

[10] In tables B1 and B2 and figure B1 of Appendix B, we provide additional robustness analyses 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.

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

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

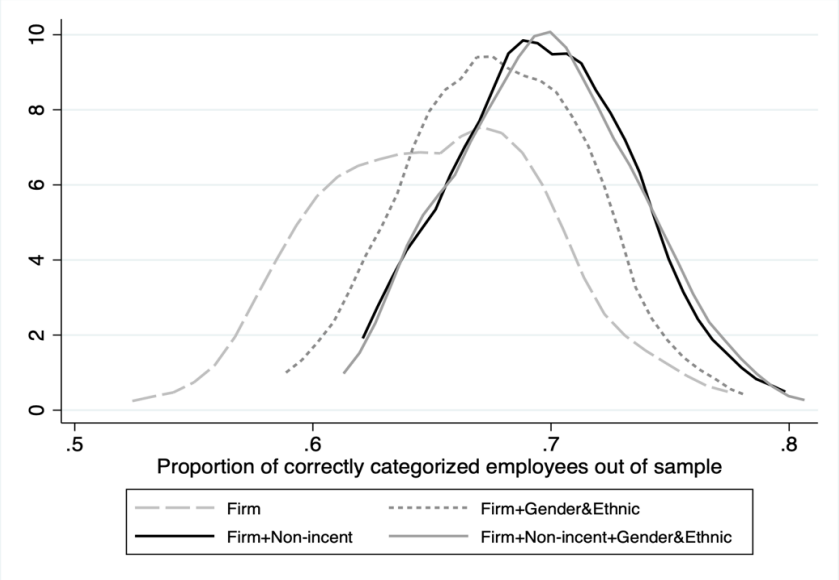


Figure 3: Validation accuracy of the random forest with and without gender and ethnicity as predictors.

[11] There are two ways to interpret this. It is possible that gender and ethnicity are the main predictors of productivity, in which case the non-incentivized measures are predictive only to the extent that they themselves help to predict gender and ethnicity. Alternatively, it might be that the main predictors are captured by the non-incentivized measures, and gender and ethnicity correlate with these measures. To distinguish between these possibilities, we train two algorithms: one predicting the propensity to receive a bonus based on the gender and ethnicity of employees only (average out-of-sample accuracy 59.5%), and the other based on the non-incentivized measures only (average out-of-sample accuracy 62.1%). The difference is significant ($p<0.01$). This suggests that the non-incentivized measures contain information beyond gender and ethnicity.

The self-reported level of risk tolerance 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 introduced by behavioral economists. The algorithm also uses some firm data, including regional effects represented by dummies for regions managed by different regional managers,[12] age, a dummy variable for holding a Bachelor's degree, a dummy for specialization in economics or management, and a dummy for being unmarried. Notably, the final algorithm includes psychological measures, particularly extraversion, neuroticism, and agreeableness that are part of the Big Five personality traits. Additionally, locus of control and a subset of responses from the Reading the Mind through the Eyes test are included. We cannot determine the direction of a variable's effect on the probability of being classified as an employee who earns a bonus, since the relationship may not be monotonous.

Variable	**Importance**
Risk tolerance 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%

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

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 tolerance 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; "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.

4 Predicting the performance of employees with short tenure

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

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

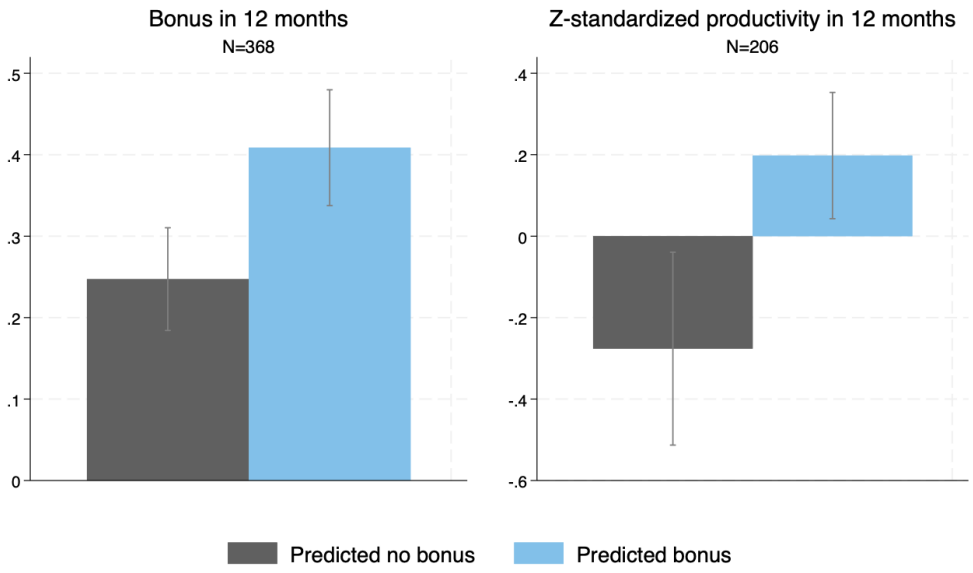


Figure 4: Bonus and productivity, conditional on employment, depending on the prediction of the algorithm.
Notes: Gray lines represent 95% confidence intervals. Sample of employees at the firm for < 12 months when responding to the survey.

Figure 4 shows outcomes by the algorithm's predictions. The left panel plots the proportion of employees who received a bonus after 12 months; the right panel displays the Z-standardized productivity measure among those still employed after 12 months. We pre-registered the measure of a bonus payment at 12 months of employment, but we also present results for the productivity as a potentially more informative measure. We did not pre-register the productivity measure since the firm did not want to commit to sharing it with us at the design stage. After deliberations with them, we received a linear transformation of the productivity measure that preserves the salary information, but allows us to analyze a cardinal measure of productivity conditional on employment. Employees predicted to receive a bonus are significantly more likely to do so than those predicted not to (two-sided Fisher's exact test, $p < 0.01$). Among employees still employed

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

at 12 months, those predicted to receive a bonus also exhibit significantly higher productivity (two-sided t-test, $p < 0.01$).

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

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

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

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

5 Design of the field experiment

The field experiment serves to evaluate the effectiveness of algorithmic hiring compared to current HR practices. Starting in March 2022, all job applicants completed a pre-interview survey hosted on the University of Lausanne's Qualtrics server. The survey only included the questions used by the algorithm based on firm data and data from the non-incentivized survey. On average, the survey took 19.5 minutes to complete. Access to the survey answers was restricted to the

researchers, and applicants were informed that their answers might be used for automated scoring. Every two to three days, we communicated the algorithm's hiring recommendations to the head office.

Two treatments were implemented:

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

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

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

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

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

Between March 2022 and February 2023, every applicant was evaluated by both the AI and HR. Overall, 1183 applicants completed the survey between March 2022 and February 2023, 590 in the AI treatment and 593 in the HR treatment. The algorithm recommended hiring 63.7% and 62.6% of applicants in treatments AI and HR, respectively (p=0.72).[15] The managers recommended hiring 96.6% and 97.8% in treatments AI and HR, respectively (p=0.22). Thus, the treatments were balanced with respect to predicted performance, but the rejection rate by the managers was well below the expected 30%. This came as a surprise to the top management, and they hypothesized that the low rejection rate might be due to a shortage of employees in 2022. For our experiment, the fact that the local and regional managers recommended to reject only very few applicants means that we have less variation across treatments than expected. Only a small number of applicants hired in the AI treatment would have been rejected by the managers. Thus, almost all treatment variation is due to 23% of the 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.[16] Finally, 536 applicants ended up being hired by the firm. Of these 536 applicants, 327 were hired in the HR treatment of which 62% were recommended by the algorithm. In addition, 209 applicants were hired in the AI treatment, 96.7% of them recommended by the managers.

6 Results of the field experiment: Algorithmic versus HR hiring

6.1 Treatment effects

As pre-registered, we examine the full set of outcomes: (i) portfolio size, (ii) number of loans issued, (iii) the portfolio with delayed repayments, (iv) probability of leaving the firm within one year, and (v) probability of receiving a bonus.[17] In addition, we obtained access to a continuous measure of *productivity*,[18] which determines the bonus amount, a figure the firm was not

[15] Note that the resulting rejection rate is higher than the target rejection rate of 30% due to differences between the pool of existing employees used to calibrate the rejection threshold and the pool of applicants.
[16] This is most likely driven by a salary offer that is lower than expected. Candidates only learn about the exact conditions of employment during the interview stage.
[17] We pre-registered those measures at 6 and 12 months of tenure. The 6 months results are in Table B3 of the appendix. In our context, productivity differences emerge with a delay due to close monitoring in the first months of employment and to lagged realized risk of loans.
[18] We have access to the Z-standardized measure of productivity which preserves secrecy of salary information.

prepared to share with us at the design stage. Note that portfolio with delayed repayments and productivity are measured conditional on continued employment after one year, while all other outcomes are observed regardless of employment status.[19]

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

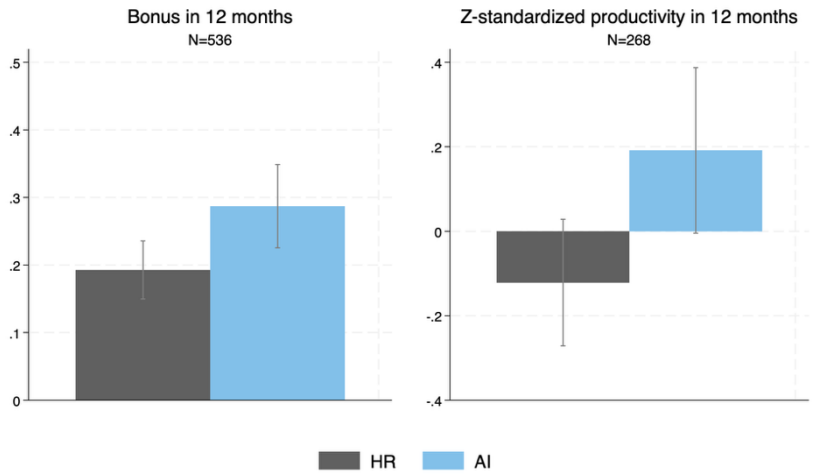


Figure 5: Treatment effect on bonus and productivity.
Notes: Gray lines represent 95% confidence intervals. Sample of all applicants hired in left panel; applicants still employed after 12 months in right panel.

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

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

[20] In the preregistration, we announced analyses based on two intervals: 6 months and 1 year. Ex post, the one-year horizon turned out to be the more relevant interval for detecting meaningful differences, as repayment problems and most performance-related issues typically materialize later in the credit cycle. Nevertheless, for completeness and to remain consistent with the preregistered plan, we report the 6-month results in the Appendix (Table B.3).

(p = 0.07 and 0.06, respectively). All other outcomes are statistically indistinguishable between the AI and HR treatments.

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

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

The mixed results regarding differences between employees hired by managers and by the AI can be due to the lack of treatment variation caused by the low rejection rate by the managers. Our findings suggest that algorithmic hiring yields a slight improvement in efficiency over current HR practices in identifying productive employees.

6.2 Profitability of AI versus HR hiring?

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

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

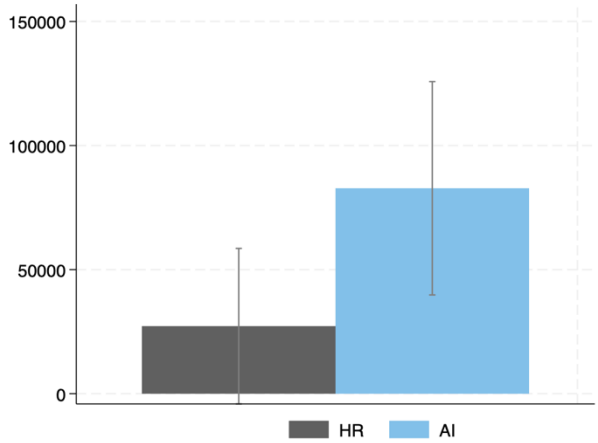


Figure 6: Treatment effect on profit per loan officer.
Notes: Gray lines represent 95% confidence intervals. Applicant sample with N=536. Profits are reported in transformed units obtained via a linear transformation of the underlying values to preserve confidentiality.

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

These findings suggest that algorithmic hiring not only improves average employee productivity but also enhances the firm's aggregate profitability, even over a relatively short time horizon. The low rejection rate by HR in the period we are considering may contribute to this result.

6.3 Potential bias under algorithmic hiring

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

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

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

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

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

The previous analysis compared the efficiency of algorithmic and HR hiring in selecting high-productivity candidates using different indicators of productivity. However, productivity alone does not define a good employee. It is crucial to consider whether the algorithm prioritizes high performance of employees at the expense of collegiality and teamwork. Relatedly, the AI may hire applicants that local managers find unsuitable for reasons beyond productivity. Notably, the bonus system is purely based on individual performance, requiring no cooperation with colleagues. Yet, office atmosphere and team spirit might influence turnover. It is therefore important to assess how the employees selected by the algorithm compare to others in key non-performance skills.

To investigate this question, we administered a survey among local managers in January 2024. The aim of the survey was to obtain an informal evaluation of employees. We analyze whether the scores for collegiality etc., obtained with the survey, are correlated with the propensity of the employees to receive a bonus and with the algorithmic recommendation. The detailed explanations of the analyses and the table with the results of the OLS regressions are in the appendix (Table B.5). We find that the scores of employees provided by local managers are consistently higher for employees who receive a bonus than for those who do not. The scores are, however, not significantly associated with the recommendation of the algorithm. We conclude that although the algorithm is designed to predict a bonus payment, it does not select applicants with significantly weaker teamwork and social skills.

7 Validating the hiring algorithm and the role of behavioral measures

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

7.1 Algorithm predicting the performance of applicants

Since the algorithm was trained to predict the probability of receiving a bonus, we first test whether employees recommended for hiring by the algorithm are more likely to receive a bonus and exhibit higher productivity than those whom the algorithm did not recommend. Among the 536 applicants hired across treatments, 62 percent were recommended for hiring by the algorithm. Figure 7 shows that employees recommended by the algorithm are significantly more likely to receive a bonus at 12 months and display higher productivity conditional on employment (two-sided Fisher exact test for the probability of receiving a bonus at 12 months and two-sided t-test for productivity both result in $p < 0.01$).

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

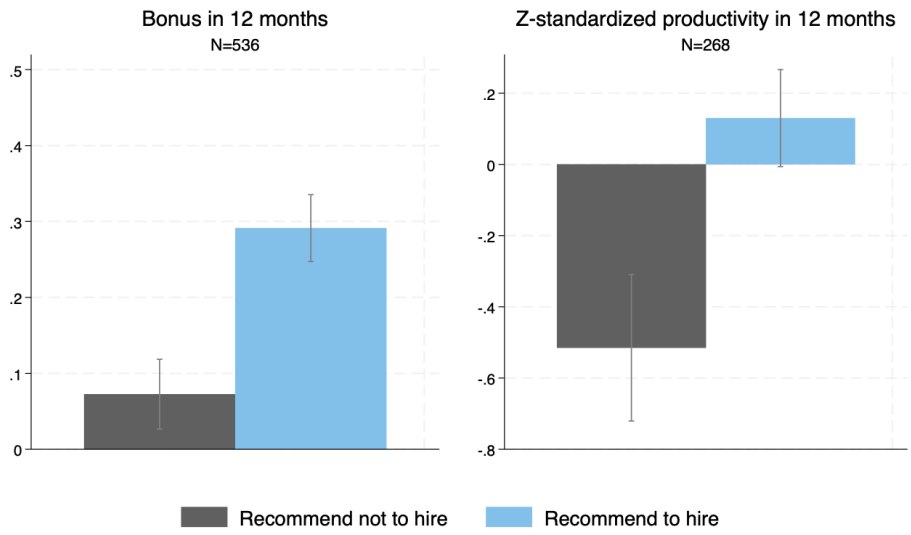


Figure 7: Proportion of employees who obtained a bonus and productivity conditional on employment, depending on the algorithm's hiring recommendation.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

We next examine all five pre-registered outcomes plus productivity. Table 5 reports the effects of the algorithmic recommendation across the full set of measures. The third row presents Romano–Wolf adjusted p-values to account for multiple hypotheses testing. Employees recommended for hiring by the algorithm perform better than those not recommended with respect to bonus payment, the share of the portfolio with delayed repayments, and productivity (significance at the 5-percent level after multiple-hypothesis testing) while difference for the remaining three measures is merely significant at the 10 percent level without corrections. These findings are not surprising given that the algorithm was trained to predict bonus payments and bonus payments are directly related to productivity.

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

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

Overall, the random forest algorithm trained on firm data and non-incentivized behavioral measures predicts applicants' future performance well. Employees recommended by the algorithm are not only more productive with a higher chance of receiving a bonus but also manage portfolios with fewer delayed repayments. These results closely mirror the findings from the analysis of existing employees with shorter tenure, except that the algorithm does not significantly predict the probability of leaving the firm within one year. This is possibly because some productive employees leave the firm because they find a better job elsewhere. Importantly, we do not observe the reason for separation—whether an employee quit or was terminated—only that the employee left the firm. We conclude that the algorithm's predictive accuracy is robust to both selection in the training data and potential strategic survey responses.[22]

7.2 Opening the black box of the algorithm's recommendation

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

Figure 8 presents the coefficient plot for the hiring recommendation, where each input of the algorithm is used as an outcome variable and regressed on a dummy of whether the algorithm recommended hiring. All inputs were z-standardized to simplify comparisons; thus, the coefficients' magnitude can be interpreted in standard deviations. The inputs are presented in order of their importance for the random forest. First, those recommended for hiring by the algorithm are significantly less risk-loving than those not recommended, i.e., they report a 0.7 standard deviation lower risk tolerance on a scale from 0 to 10. Also, they report 0.5 standard deviations more patience on a scale from 0 to 10 and 0.6 standard deviations more confidence in the number of correct answers for the CRT and numeric literacy tests. We also observe that those recommended for hiring are 0.25 standard deviations more numerically literate, score lower on the neuroticism scale, and are more likely to have a bachelor's degree compared to those not recommended. All other inputs do not vary significantly between those who are recommended for hiring and those who are not, despite their importance for the algorithm. This means that the effects of these inputs are not monotonous. For instance, age is the fifth most important variable

[22] Note that the results become even stronger if we use continuous instead of binary predictions by the algorithm. The analysis is presented in Appendix B and summarized in Table B.4.

for the algorithm, but the average age does not differ between those recommended to be hired and those who are not.

7.3 Are responses strategic?

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

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

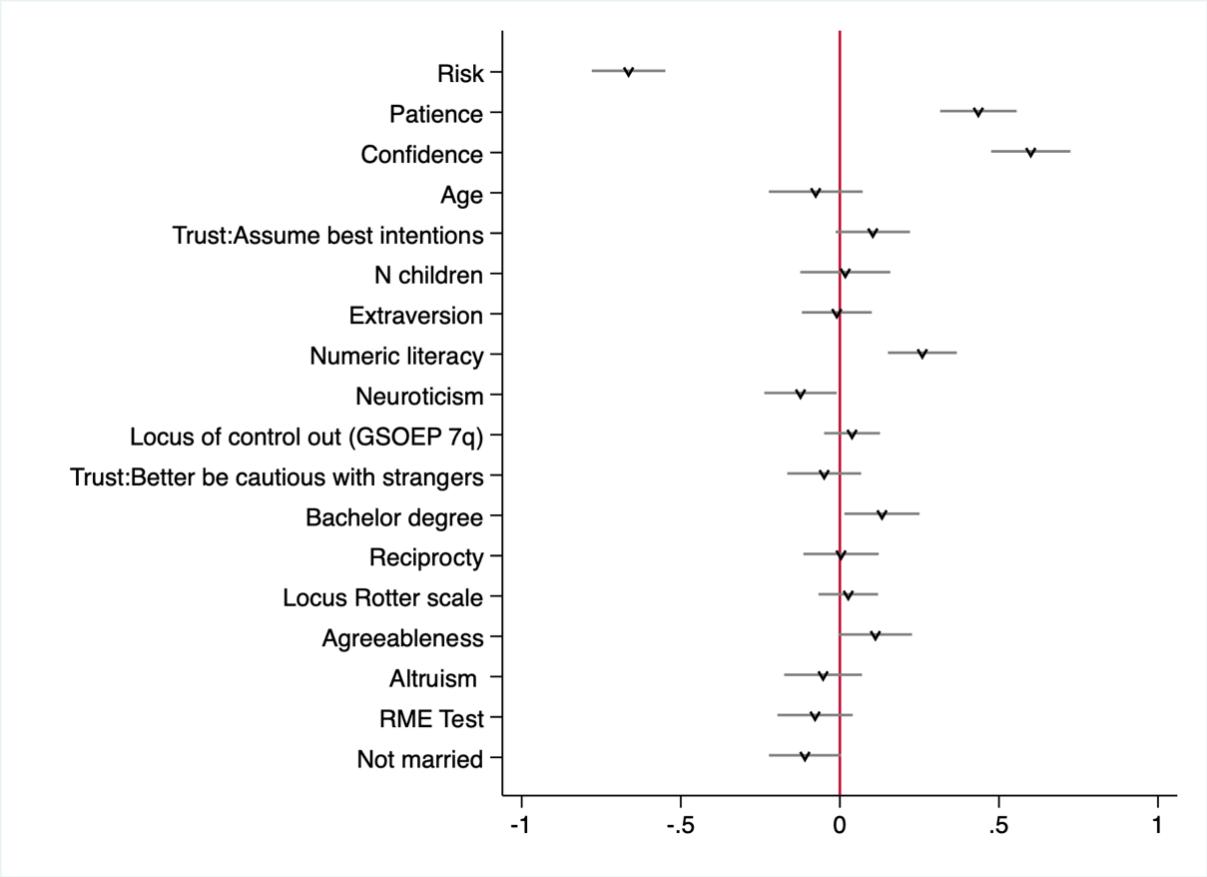


Figure 8. Coefficient of predicting recommendation to hire in OLS with the measure as outcome.
Notes: The red vertical line references zero. See notes of Table 1 for description of each measure.

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.

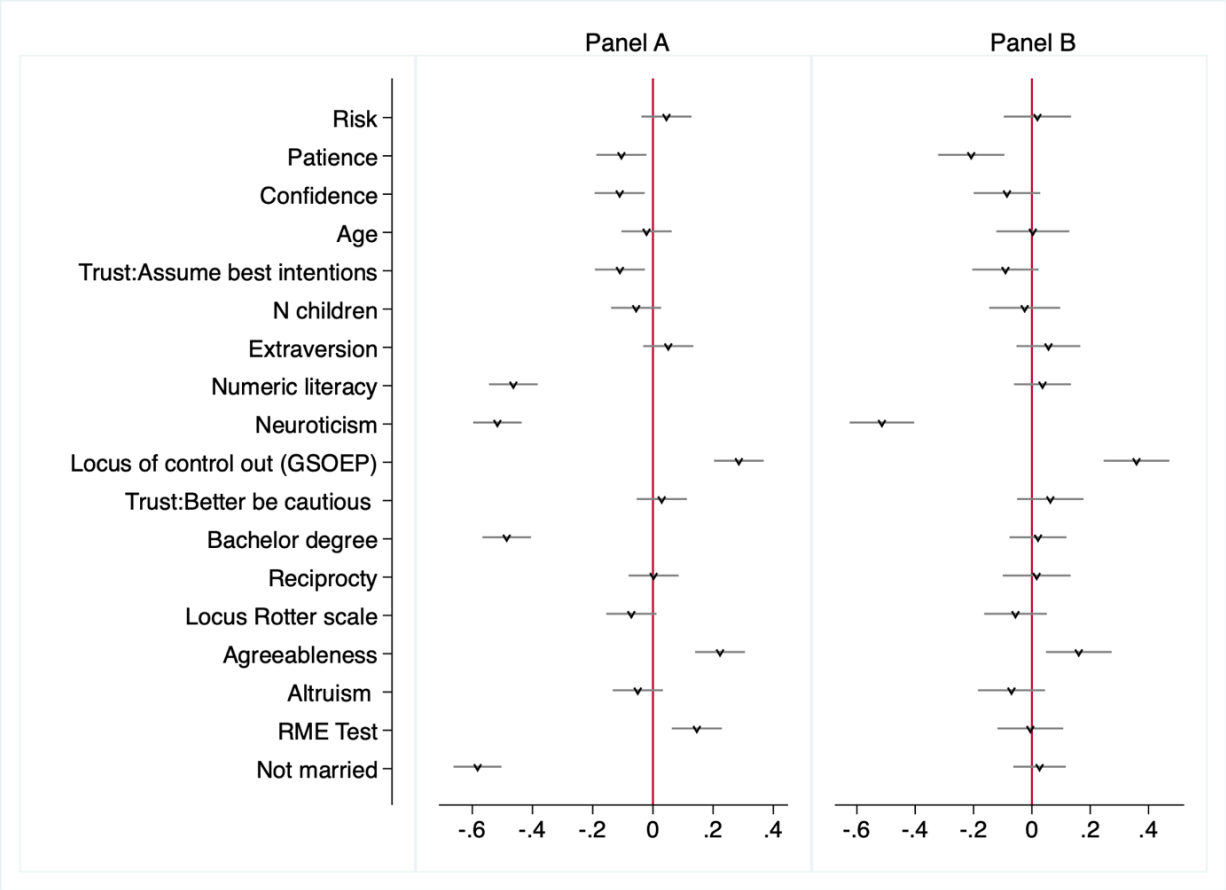


Figure 9. Coefficient plots of the OLS regression of the input variables of the algorithm on the dummy for the applicant sample relative to the employee sample.

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[23], and the RME test. We selected these variables because the answers to them are likely to be non-strategic, assuming that applicants answer the questions of the CRT, numeric literacy, and RME test to the best of

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

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 differ between the employee and applicant samples than without the controls. None of the non-strategic variables significantly differ, as expected when controlling for propensity scores. However, four inputs to the algorithm remain significantly different between the two samples. First, applicants report lower levels of patience compared to employees. This result is surprising as patience is valued by the algorithm (see Figure 8) and applicants should - if anything - 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 to improve their hiring chances. A lower score on neuroticism is expected to increase the likelihood of receiving a hiring recommendation, as shown in the previous subsection. For locus of control and agreeableness, however, the overall effect is zero, since it affects the likelihood of being recommended for hiring in a non-monotonous 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 and indicate that economic measures are harder to manipulate, since their connection to productivity or employer preferences seems less predictable for candidates.

One might argue that manipulations are initially challenging, but candidates can learn to game the system. If this were the case, the proportion of candidates rejected by the algorithm should decrease over time. However, we see no indication of this during our experiment that lasted for one year. There is no upward trend in the proportion of candidates recommended for hiring by the algorithm each month, see Figure B2 in Appendix B. Additionally, none of the monthly fixed effects are significant (the lowest p-value being $p=0.22$ for June 2022 relative to March 2022). Thus, at least for the duration of our experiment, we do not observe any significant pattern of candidates adapting to the algorithm.

7.4 What is the role of behavioral measures?

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

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

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

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

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
Behavioral algorithm recommends hire	3.6e+05	13.975	0.262***	-0.057	-0.011**	0.620***
	(3.6e+05)	(12.273)	(0.056)	(0.056)	(0.004)	(0.164)
Firm algorithm recommends hire	5.3e+05	22.190	0.016	-0.072	0.002	0.067
	(4.1e+05)	(13.785)	(0.058)	(0.063)	(0.005)	(0.189)
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

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

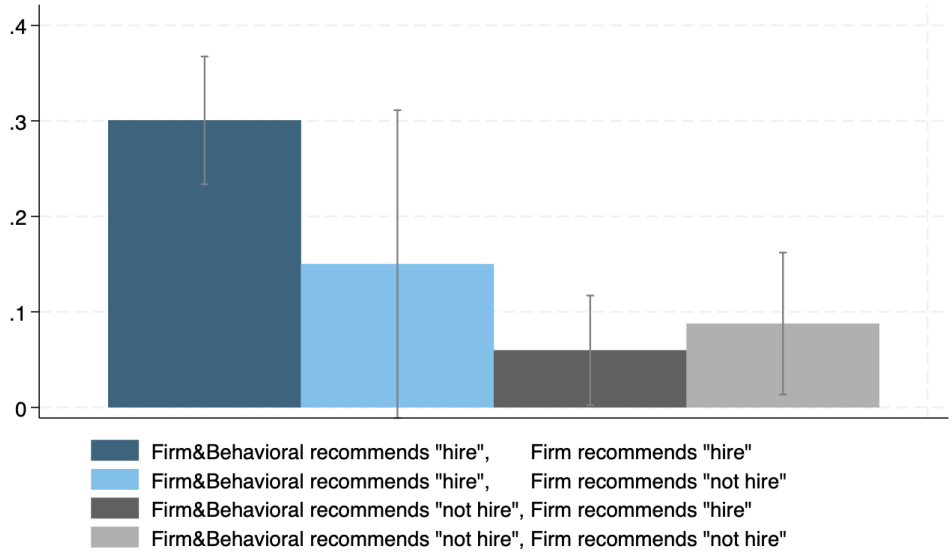


Figure 10: Proportion of employees who obtained a bonus depending on the recommendation of the two algorithms.
Notes: Gray lines represent 95% confidence intervals. Applicant sample.

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

significant. Based on these estimates, the average profit per new hire is 13.4 percent higher when using the algorithm with behavioral measures compared to the version trained only on firm data.

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

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

8 Discussion and conclusions

Our findings offer a positive perspective on the use of behavioral measures in hiring processes, thereby pointing to a new application for behavioral economics: providing standardized measures of human behavior to improve algorithmic hiring and potentially other AI applications, such as credit scoring. In hiring contexts where applicants have few observable qualities to signal their potential and where the job is such that an ideal candidate profile cannot be defined easily, survey measures enable firms to conduct more effective automated screening. It is encouraging that the algorithmic predictions are robust to selective training samples and strategic responses of candidates. While we observe some strategic behavior, particularly for personality measures such as neuroticism, agreeableness, and locus of control, the economic measures show only a small variance between the training sample and the pool of candidates, allowing the algorithm to accurately predict the employees' future performance. One possible concern is that applicants can learn over time how to answer the survey questions used for algorithmic hiring. However, many variables enter in a non-monotonous 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 study was conducted in the specific context of a microcredit firm in Kyrgyzstan. It is uncertain how well our approach performs in other environments. However, the usefulness of behavioral measures and their relative robustness to strategic manipulation can be expected to extend to other recruitment contexts. The tasks of our loan officers involve agency and balance between short-term and longer-term incentives which we believe applies to many white collar

jobs and thus, behavioral features are useful signals of productivity. Although surveys have been utilized by HR for decades, we are unaware of causal evidence that selection based on survey measures outperforms traditional methods, such as interviews. Thus, our results can be interpreted as validating and refining the technique that Daniel Kahneman introduced in the Israeli army in 1955 to improve its hiring practices. He recommended relying on a set of predetermined tests instead of forming intuitive judgments based on interviews. We add the selection and weighting of these traits with the help of AI and demonstrate its robustness to strategic responses and a biased training sample.

References

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

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

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

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

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

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

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

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

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

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

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

Bó, I., Chen, L., & Hakimov, R. (2023). Strategic Responses to Personalized Pricing and Demand for Privacy: An Experiment. arXiv preprint arXiv:2304.11415.

Bogen, M., Rieke, A. (2018): Help Wanted: An Exploration of Hiring Algorithms, Equity and Bias. Tech. rep., Upturn (2018):

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

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

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

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

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

Chapman, J., Ortoleva, P., Snowberg, E., Yariv, L., & Camerer, C. (2025). *Reassessing qualitative self-assessments and experimental validation* (No. w33520). National Bureau of Economic Research.

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

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

Dargnies, M. P., Hakimov, R., & Kübler, D. (202). Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence. *Management Science,* 72(1).

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

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

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

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

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

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.Gosnell, G. K., List, J. A., & Metcalfe, R. D. (2020). The Impact of Management Practices on Employee Productivity: A Field Experiment with Airline Captains. *Journal of Political Economy*, 128(4), 1195–1233.

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

Guenzel, M., Kogan, S., Niessner, M. & Shue, K. (2026). AI Personality Extraction from Faces: Labor Market Implications (January 30, 2026). https://ssrn.com/abstract=5089827

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

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

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

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

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

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

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

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

Hoffman, M., & Stanton, C. T. (2024). People, Practices, and Productivity: A Review of New Advances in Personnel Economics.

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

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

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.

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

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

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

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

Li, D., L. Raymond and P. Bergman (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.

Rhea, A.K., Markey, K., D'Arinzo, L. *et al.* (2022). An external stability audit framework to test the validity of personality prediction in AI hiring. *Data Min Knowl Disc* 36, 2153–2193. https://doi.org/10.1007/s10618-022-00861-0

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?

<u>Non-incentivized</u>

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?[24]

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
- 20 000
- 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.

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

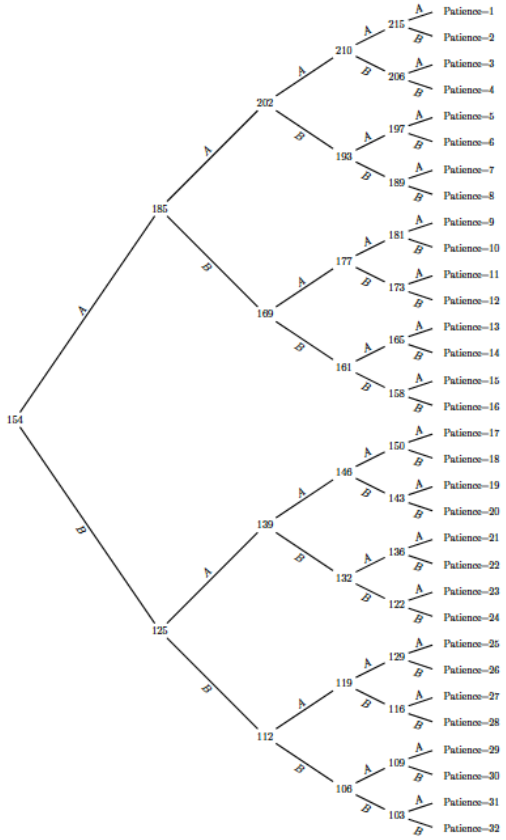


Figure 3: Tree for the staircase time task (numbers = payment in 12 months, A = choice of "100 euros today", B = choice of "x euros in 12 months"

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?

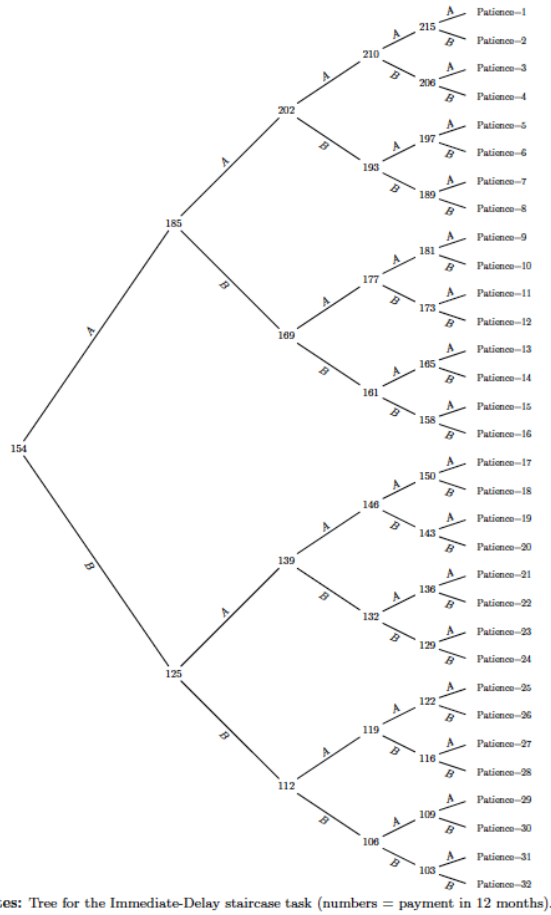
Use a scale from 0 to 10, where 0 means "completely unwilling to do so" and 10 means " completely willing to do so" (Becker):

- 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

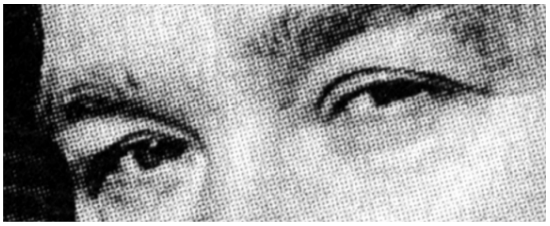
Annoyed

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 tables and figures

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%

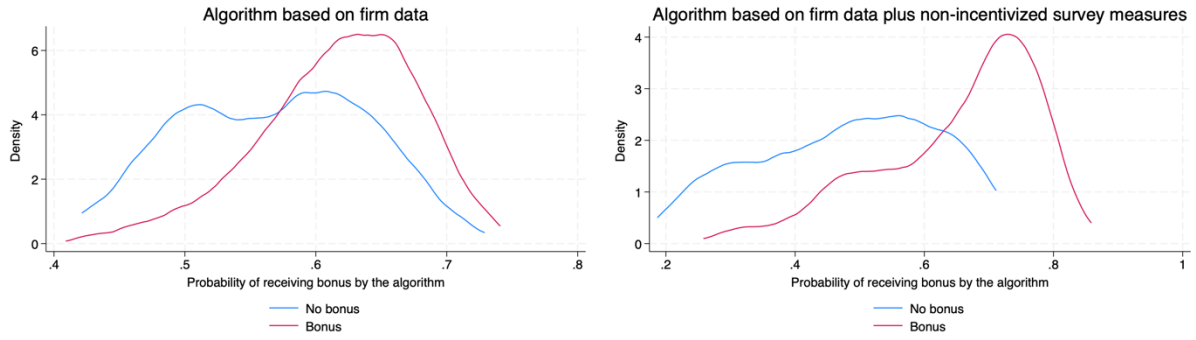


Figure B1. The density of the predicted probabilities conditional on having received a bonus or not. Left panel: The algorithm based on firm data alone. Right panel. The algorithm based on firm and non-incentivized data. To generate these graphs, we ran each model 250 times. Each

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 calculate the average probability for each employee based on all instances they were included in the validation group.

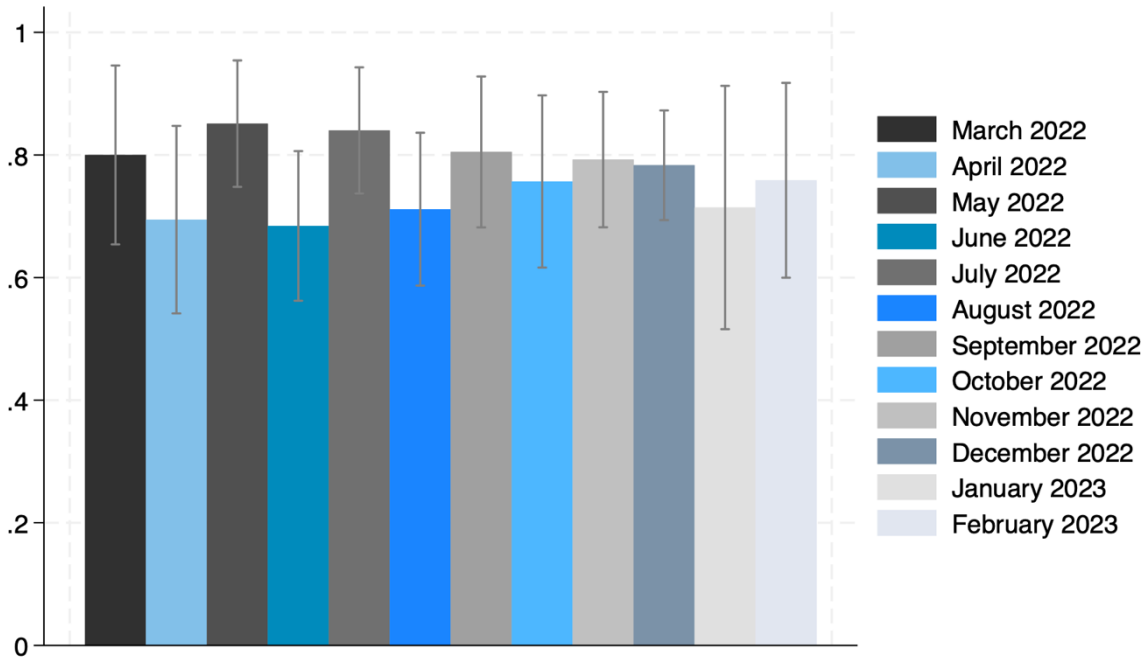


Figure B2. Proportion of candidates recommended for hiring by algorithm.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
AI treatment	3.1e+05	4.076	0.064**	-0.006	-1.4e+03	0.227**
	(2.9e+05)	(4.330)	(0.026)	(0.041)	(9918.822)	(0.111)

Romano-Wolf p-value	0.68	0.70	0.07	0.99	0.99	0.15
Observations	536	536	536	536	337	337
Sample	All	All	All	All	Employed	Employed

Table B.3: Performance of employees in 6 months. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays, and productivity, and marginal effect of probit regressions of the bonus and left-the-firm dummies on the AI treatment dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
Algorithm recommends hire	5.0e+05	21.446*	0.263***	-0.106*	-8.7e+04***	0.587***
	(3.7e+05)	(12.401)	(0.055)	(0.056)	(2.0e+04)	(0.161)
AI treatment	1.2e+05	1.771	0.009	0.044	1.9e+04	0.119
	(3.2e+05)	(10.722)	(0.039)	(0.049)	(1.7e+04)	(0.132)
Romano-Wolf p-value	0.15	0.10	0.003	0.10	0.003	0.003
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

Table B.4 : Performance of employees in 12 months controlling for the treatment assignment. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays, and productivity, and marginal effects of probit regressions of the bonus and left-the-firm dummies on the algorithm recommendation dummy. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Appendix C. Additional results

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 only 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.4 presents results analogous to Table 3 using the continuous variable of the percentage of hiring recommendations by the algorithm instead of the binary measure.

	Portfolio size	Issued loans	Bonus	Left the firm	Portfolio with delays	Productivity
% of runs where algorithm recommends hiring	3025.9**	0.117***	0.001***	-0.000*	-300.5***	0.003***
	(1309.382)	(0.044)	(0.000)	(0.000)	(72.909)	(0.001)
Romano-Wolf p-value	0.04	0.02	0.003	0.07	0.003	0.003
Observations	536	536	536	536	268	268
Sample	All	All	All	All	Employed	Employed

Table B.4: Performance of employees. Coefficients of OLS regression for portfolio size, issued loans, portfolio with delays, and productivity, and marginal effects of probit regressions of the bonus and left-the-firm dummies 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 and coefficients of the OLS regressions reported in Table B.4 are generally stronger and more precisely estimated than in Table 5This difference reflects the fact that Table B.4 exploits variation in the proportion of algorithmic runs recommending hiring, rather than a binary indicator based on a fixed cutoff calibrated to HR selectivity. The share of positive recommendations therefore provides cardinal information about applicant quality that is lost when recommendations are dichotomized. In turn, this suggests that hiring decisions need not rely on a single threshold rule: instead, the cutoff used to translate algorithmic scores into hiring decisions could be adjusted or optimized to better align with predicted job performance.

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.5 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	0.739*** (0.262)	1.346*** (0.302)	0.458* (0.240)	10.870* (5.898)	0.787*** (0.238)
Algorithm recommended hire	-0.285 (0.350)	-0.525 (0.364)	0.026 (0.373)	-7.600 (7.515)	-0.019 (0.312)
Constant	8.096*** (0.260)	7.322*** (0.326)	8.412*** (0.364)	71.724*** (6.204)	1.683*** (0.263)
Observations	204	204	204	204	204
R^2	0.039	0.107	0.018	0.020	0.073
Sample	Employed 02.24	Employed 02.24	Employed 02.24	Employed 02.24	Employed 02.24

Table B.5. 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.5 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.[25] 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.

[25] 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.