Task Complexity and Firms' Willingness to Replace Entry-Level Labour with AI: A Survey Experiment with Managers in Germany

Vincent Rost [*1,2,3] and Fabian Stephany[3,4,5]

[1] *ZEW Mannheim, Germany*
[2] *Justus Liebig University Giessen, Germany*
[3] *Oxford Internet Institute, United Kingdom*
[4] *Institute for New Economic Thinking, United Kingdom*
[5] *Bruegel, Belgium*

October 14, 2026

Abstract

How does task complexity shape firms' willingness to substitute entry-level labour with artificial intelligence (AI) tools? We study this question using a pre-registered survey experiment with managers at 1,435 German firms in the information economy and manufacturing industry. Managers were randomly assigned to consider a persistent need for additional capacity involving either standardised tasks or complex tasks requiring professional judgement, contextual knowledge, and independent problem-solving. Relative to standardised tasks, complex tasks reduce willingness to use AI rather than hire an entry-level employee by 0.54 points on a five-point scale and increase the importance attached to human oversight by 0.15 points. Pre-treatment responses further indicate that managers expect AI availability to reduce hiring of both entry-level and experienced workers, but significantly more strongly for entry-level workers.

The work was made possible with the support of a scholarship from the German Academic Exchange Service (DAAD).

Keywords: AI, labour market, survey experiment, managerial decisions, hiring decisions

[*] Corresponding Author; Vincent.Rost@zew.de.

1 Introduction

The diffusion of artificial intelligence (AI) and generative AI in particular has renewed concerns about technological displacement and employment. Yet, the emerging debate increasingly suggests that AI's consequences are unlikely to be uniform across the workforce. Rather than simply asking whether AI creates or destroys jobs, a more consequential question is which workers are most exposed to substitution, and why. One particular concern is that AI can disproportionately affect workers at the beginning of their careers.

Recent evidence provides some support for this possibility, but remains contested. Employment and hiring have declined among young workers in highly AI-exposed occupations in several settings (Brynjolfsson et al., 2026; Lodefalk et al., 2026; Tucker, 2026). Brynjolfsson et al. (2026), for example, find that employment among 22–25-year-olds in highly AI-exposed US occupations has fallen substantially relative to less-exposed peers, while experienced workers show no comparable decline. However, other studies find weaker evidence once concurrent labour-market developments are considered. Lambert and Schindler (2026) show that AI exposure is strongly correlated with remote-work exposure and that the association between AI and declining junior hiring attenuates considerably when both are considered jointly, while Fairlie and Wu (2026) find no significant relative increase in unemployment among recent US college graduates. Thus, whether AI systematically reduces demand for junior workers and, importantly, why it might do so, remains unresolved.

Several mechanisms could generate such an effect. AI may disproportionately increase the productivity of junior workers, allowing firms to produce the same output with fewer employees (Brynjolfsson et al., 2025). Alternatively, AI may substitute for the relatively standardised tasks disproportionately assigned to junior workers, consistent with task-based theories of technological change (Acemoglu and Restrepo, 2019; Autor, 2013; Hampole et al., 2025). AI may also lead to a seniorisation of entry-level work, as automation of simpler activities leaves junior workers responsible for tasks requiring greater judgement, problem solving, and communication (Westby et al., 2026). Finally, AI may alter organisational complementarities by allowing experienced employees to supervise AI rather than teams of junior workers. These mechanisms imply very different transformations of entry-level work and career progression.

In this paper, we focus on one central dimension that can help distinguish between these explanations: task complexity. We conduct a preregistered survey experiment with 1,435 managers from German

firms in the information economy and manufacturing. Managers are randomly assigned to consider a persistent need for additional capacity involving either standardised, clearly structured entry-level tasks or complex entry-level tasks requiring professional judgment, contextual knowledge, and independent problem-solving. We then ask whether they would use AI instead of hiring an additional entry-level employee and how important they consider human oversight of AI-generated output.

We find that managers' willingness to substitute AI for entry-level labour depends strongly on task content. For complex rather than standardised tasks, willingness to use AI instead of an additional entry-level employee falls by approximately 0.54 points on a five-point scale. At the same time, task complexity significantly increases the importance managers attach to human oversight of AI-generated outputs. Complementary evidence on hiring expectations reveals that managers anticipate AI reducing hiring of both junior and experienced workers, but significantly more strongly for entry-level employees. Thus, the expected adjustment is not an absolute expansion of experienced employment but a relative shift in labour demand towards experience.

Our findings contribute to the emerging literature by moving beyond whether junior employment is declining toward the organisational mechanism through which AI may affect it. Junior status alone does not determine substitutability; what junior workers do matters. This complements evidence that AI simultaneously substitutes for some human capabilities while increasing demand for complementary ones (Mäkelä et al., 2026), and that the composition and complexity of workers' skills shape their resilience to technological change (Carpanelli et al., 2026).

More broadly, our findings suggest that AI may disproportionately automate standardised work through which workers have traditionally entered occupations, while preserving human involvement where work requires judgment, contextual knowledge, and oversight. This raises a longer-term question beyond immediate employment effects: if AI substitutes for the tasks through which junior workers traditionally acquire expertise, how will firms develop the experienced workers they continue to value?

The remainder of the paper develops the competing mechanisms and hypotheses, describes the survey experiment and empirical strategy, presents the experimental and hiring-expectation results and robustness analyses, and concludes by discussing their implications for entry-level work and career progression.

2 Conceptual Background and Hypotheses

AI does not substitute for human capabilities uniformly. Its diffusion can simultaneously reduce demand for substitutable skills while increasing demand for complementary ones (Mäkelä et al., 2026). One prominent concern is that these effects may disproportionately affect workers at the beginning of their careers. Recent studies document declining employment or hiring among young workers in highly AI-exposed occupations in the United States and Sweden (Brynjolfsson et al., 2026; Lodefalk et al., 2026; Tucker, 2026). Yet the evidence remains contested. The apparent AI effect becomes considerably weaker when accounting for remote work in some settings (Lambert and Schindler, 2026), while other studies find no distinct post-LLM decline in junior hiring or unemployment (Fairlie and Wu, 2026; OECD, 2026). Thus, both the extent to which AI reduces demand for junior workers and, importantly, why it might do so remain unresolved. We distinguish four mechanisms that can generate declining demand for junior workers while implying fundamentally different transformations of work.

2.1 Productivity Paradox

First, a productivity paradox may arise if AI disproportionately increases junior-worker productivity. Generative AI produces substantial productivity gains in knowledge work (Dell'Acqua et al., 2023; Noy and Zhang, 2023), with some of the largest gains accruing to less experienced workers (Brynjolfsson et al., 2025). Yet greater individual productivity need not increase labour demand. Unless productivity gains generate sufficiently large scale effects, firms may require fewer junior employees to produce the same output. AI can therefore complement individual junior workers while simultaneously reducing firms' demand for them (Acemoglu and Restrepo, 2019; Hampole et al., 2025).

2.2 Task Substitution

Second, AI may produce task substitution. Technologies substitute for tasks rather than occupations wholesale (Acemoglu and Restrepo, 2019; Autor, 2013), and many tasks conventionally assigned to junior workers, such as information retrieval, drafting, summarisation, basic analysis, translation, or routine coding, overlap with current AI capabilities. Greater AI exposure across an occupation's task bundle has accordingly been associated with declining labour demand (Hampole et al., 2025), while AI diffusion reduces demand for potentially substitutable human skills and increases demand for complementary

capabilities (Mäkelä et al., 2026). Junior workers may therefore be vulnerable not because they are junior per se, but because of the tasks they perform. Given AI's "jagged technological frontier" (Dell'Acqua et al., 2023), this mechanism predicts considerably weaker substitution where junior workers perform more complex tasks.

2.3 Seniorisation

Third, AI may lead to a seniorisation of junior work. Rather than eliminating entry-level positions, firms may shift their content toward tasks requiring problem solving, judgment, communication, and evaluation. Westby et al. (2026), for example, find declining junior relative to senior software vacancies after the introduction of ChatGPT alongside rising experience requirements and greater emphasis on problem solving and communication within remaining junior roles. Similarly, AI-intensive jobs increasingly reward complementary human capabilities such as analytical thinking and resilience (Mäkelä et al., 2026). AI may therefore raise the threshold for what constitutes an entry-level worker, increasing demand for junior employees who already possess capabilities conventionally associated with more senior roles.

2.4 AI Complementarity

Fourth, AI may create a new senior–AI complementarity. Knowledge-intensive production has traditionally combined junior workers executing tasks with senior workers providing expertise, judgment, and supervision. Generative AI potentially alters this relationship by allowing experienced workers to delegate execution to AI and supervise its outputs instead. The relevant production complementarity may consequently shift from senior–junior toward senior–AI. This is consistent with evidence that automating some tasks can increase the value of human input into remaining tasks (Hampole et al., 2025), particularly capabilities complementary to AI (Mäkelä et al., 2026), and that human expertise becomes particularly important when tasks extend beyond AI's capabilities (Dell'Acqua et al., 2023).

2.5 Investigating Task Complexity

Distinguishing these mechanisms requires moving beyond occupational titles and seniority toward the tasks and capabilities underlying work. Recent evidence demonstrates that the structure of workers' skill portfolios independently predicts career outcomes and resilience: specialisation is associated with

productive depth, diversity with mobility, and workers combining both are particularly likely to move toward occupations with lower automation exposure (Carpanelli et al., 2026). Task complexity may therefore provide an important empirical means of distinguishing mechanisms that otherwise produce observationally similar declines in junior employment.

We test these mechanisms using a firm survey of German companies in the information and manufacturing sectors, distinguishing workers by both seniority and task complexity. The mechanisms generate different empirical predictions. A productivity mechanism predicts reduced demand for junior workers irrespective of task complexity, because the relevant distinction is worker experience. Task substitution instead predicts that reduced junior demand should be concentrated among simple rather than complex tasks. Seniorisation predicts greater demand for junior workers possessing skills and capabilities traditionally associated with senior roles. Finally, senior–AI complementarity predicts greater importance of human supervision as task complexity increases and, specifically, stronger demand for senior workers to provide such supervision.

Together, these mechanisms shift the question from whether AI simply replaces junior workers toward how AI reorganises tasks, expertise, and supervision across levels of seniority. We hypothesise that AI's effect on demand for junior workers is conditional on the nature and complexity of their tasks, with differences across task complexity and seniority revealing the mechanisms through which AI reshapes entry-level work.

2.6 Hypotheses

This implies that the substitutability of entry-level work depends on the content and the nature of the task for which the firm requires additional capacity. Standardised and clearly structured tasks can more readily be specified, delegated to an AI system, and assessed against a pre-defined output. When tasks require contextual knowledge, professional judgement, and independent problem-solving, however, the task is harder to delegate to AI tools. We therefore expect managers to be less willing to replace an additional entry-level worker with AI when the required work is complex:

H1: Firms are more likely to use AI applications instead of hiring an additional entry-level employee when the required tasks are standardised and clearly structured rather than complex.

Task complexity may also affect how firms intend to use AI when they do deploy it. Outputs from

standardised processes are comparatively easy to verify, while errors in context-dependent tasks may be harder to detect and have more severe consequences. We therefore assume that complex work should increase the perceived importance of human oversight of potential AI outputs.

H2: Firms attach greater importance to human oversight of AI-generated output when the required tasks are complex rather than standardised and clearly structured.

Hypotheses 1 and 2 concern different organisational decisions. The first addresses whether AI is considered an alternative to an additional entry-level hire. The second concerns the degree of human involvement managers consider necessary when AI is used.

Beyond the experimental scenario – i.e. before the experimental treatment takes place – the survey elicits managers' expectations about how the availability of AI tools will affect their firm's hiring decisions over the following two years. Entry-level positions often contain standardised or codifiable activities that can be at least partly automated. If AI allows existing employees to complete these activities more efficiently, firms may require fewer additional junior workers. This leads to the following prediction:

H3: Managers expect the availability of AI applications to reduce their firm's hiring of entry-level employees over the following two years.

The expected effect on experienced employees is less clear. AI may automate some of their activities as well, but at the same time, expertise and domain knowledge will be needed to oversee and critically judge AI outputs. Our preregistered hypothesis therefore emphasises complementarity between AI applications and experienced[1] workers:

H4: Managers expect the availability of AI applications to increase their firm's hiring of experienced employees over the following two years.

Hypotheses 3 and 4 therefore predict that AI serves as a substitute for junior-level workers, but as a complement for experienced workers. They are evaluated separately against the neutral response category of no expected effect. Unlike Hypotheses 1 and 2, these comparisons are descriptive. The questions are causal in their wording, as managers are directly asked how the availability of AI will affect their firm's future hiring. This causal judgement is, however, supplied by the respondent, rather than identified through experimental variation. The estimates therefore capture managers' perceived causal effects of

[1] Our questionnaire defined those as "employees with professional experience, i.e., those with in-depth knowledge in their field"

AI availability, not the causal effect of AI on realised employment. We additionally compare the two expected hiring effects within firms.

These four hypotheses constitute the confirmatory hypotheses specified in the preregistration.[2]

3 Data and Experimental Design

3.1 Survey and Sample

The firm survey 'Konjunkturumfrage Informationswirtschaft' that is conducted every quarter by ZEW (Leibniz Centre for European Economic Research, Mannheim, Germany) is the data source for this research project. In its current format, it has been running since 2011 and produces between 1,000 and 1,500 responses each quarter. Originally, this survey only covered firms that belong to the information economy (e.g. ICT hardware, ICT services and knowledge-intensive services such as media agencies, tax consultants etc.) and have at least 5 employees. Recently, however, the scope of the survey was extended as it now also regularly includes firms in the manufacturing industry (also with at least 5 employees), which also allows comparisons across both industries.

The questionnaires are sent by post directly to the CEOs of the firms. Even though we can't fully rule out that the paper questionnaire is handled by other employees within the company, we are fairly confident that the vast majority of our respondents has substantial decision power in the firm. In a previous wave of the panel we explicitly asked the respondent for their position within the firm: Around 73% of the respondents were CEOs or members of the executive board, 17% reported to be department heads (or holding higher managing positions), while less than 3 percent were team or project leaders. Only 7 percent of respondents did not hold any management function. Even though these numbers stem from another wave of the survey, we are confident to indeed reach CEOs and decision-makers of German firms with our instrument.

After receiving the questionnaire, which consists of only one page with easy-to-answer items, they are presented with several ways to respond: They can either use the provided paper questionnaire and send the filled-in version back to ZEW via post or scan it and send it via e-mail or fax machine. Alternatively, they can use an online log-in to answer the questions using an online form. In total, 1,059 respondents

[2]The preregistration is available at `https://aspredicted.org/ix9za2.pdf`.

(73.8 per cent) participated online and 376 (26.2 per cent) used an offline mode.

The data for this paper come from wave 129 of the aforementioned firm survey, which was fielded in mid 2026. Online responses were recorded between 19 June and 5 August 2026. A small number of paper questionnaires received after the online field period were also retained, with the final accepted questionnaire entered on 26 August 2026. Alongside recurring questions on business conditions and expectations, each wave contains modules on digitalisation, innovation, and the organisation of work. This wave focused on firms' use of artificial intelligence, hiring behaviour and their demand for entry-level labour.

After excluding firms that do not belong to the pre-defined industry or have fewer than five employees, the final sample consists of 1,435 observations. Of these, 768 (53.5%) belong to the information economy, 667 (46.5%) operate in the manufacturing industry. Firms are distributed fairly evenly across the three size classes used in the analysis:[3] 448 (31.2%) have 5–19 employees, 550 (38.3%) have 20–99 employees, and 437 (30.5%) have at least 100 employees. The median firm has 38 employees, whereas the mean is 172.9.

Regarding personal characteristics, among respondents identifying as either female or male, 24.7% are female and 75.3% are male.[4] As the survey directly addresses managers and employees with high levels of decision power, the age distribution is shifted towards the upper end of the scale: 71.3% of respondents are at least 50 years old.[5]

Interestingly, the sample was already relatively familiar with AI applications. The firms report using an average of 2.2 of the six AI application categories listed in the questionnaire, and 67.3 per cent of respondents use AI applications either daily or several times per week. The experiment therefore elicits responses and hiring expectations from managers with often direct exposure to AI. Table 3.1 reports the composition of the sample along categorical (Panel A) and numerical (Panel B) characteristics.[6]

[3]This is not representative for the size distribution of German firms; pooled results should therefore be treated with caution, as large firms are overrepresented in our sample

[4]Two respondents selected "other" as their gender, 49 selected "prefer not to answer", and 46 did not provide a gender response.

[5]Forty respondents did not report their age category.

[6]The analytical samples for the outcomes are documented in Appendix Table B.6.

Characteristic	Category	N	Percent	Mean	SD	Median
Panel A: Sample composition						
Industry	ICT sector	216	15.1			
	Media services	80	5.6			
	Knowledge-intensive services	472	32.9			
	Manufacturing	667	46.5			
Firm size	5–19 employees	448	31.2			
	20–99 employees	550	38.3			
	100+ employees	437	30.5			
Survey mode	Offline	376	26.2			
	Online	1059	73.8			
Respondent characteristics	Female	330	24.7			
	Aged 50 or older	995	71.3			
AI experience	Frequent personal AI use	966	67.3			
Previous entry-level hiring	Vocational qualification	760	53.0			
	University degree	655	45.6			
Panel B: Numeric characteristics						
Employees in 2025		1403		172.9	879.5	38.0
Employees with a university degree (%)		1316		31.5	31.1	20.0
AI applications used		1430		2.2	1.5	2.0
Hiring difficulty: vocational		1107		3.0	1.3	3.0
Hiring difficulty: university		1039		2.9	1.3	3.0

Table 3.1: Sample characteristics

Notes: The table describes the eligible sample of 1,435 firms. Except for respondent gender and age, percentages in Panel A are calculated relative to the full eligible sample. The share of female respondents is calculated among respondents who identified as either female or male ($N = 1{,}338$); two respondents selected another gender category, 49 preferred not to answer, and 46 did not provide a gender response. The share of respondents aged 50 or older is calculated among respondents reporting a valid age category ($N = 1{,}395$); 40 respondents did not report their age. For binary characteristics, N reports the number of observations meeting the stated condition; the remaining observations belong to the respective reference category or, where applicable, have missing information. Frequent personal AI use is defined as using AI tools daily or several times per week. Previous entry-level hiring indicates that the firm hired at least one entry-level employee with the respective qualification during the previous two years. The number of AI applications ranges from zero to six. Hiring difficulties are measured on a five-point scale, with higher values indicating greater difficulty, and are reported only for firms indicating a need for the respective qualification group.

3.2 Experimental Design and Randomisation

In order to causally identify the effect of task complexity on our outcome measures, each firm is randomly assigned to one of two groups. Both groups received largely identical questionnaires that only differed in the task complexity for the given scenario. Respondents were asked to imagine a scenario in which they needed additional capacities to fulfil a task and that this additional need would persist long-term. The experimental variation alters the level of complexity for the given task. In the high-complexity group, the relevant task was described as being suitable for entry-level workers, "but require professional judgement, contextual knowledge and independent problem-solving". Additionally, we provided examples to underline the type of tasks we were referring to.

The other group received a slightly different setting. The scenario was identical (long-term need for additional capacity), but the tasks in question were less complex. They were defined as suitable for entry-level workers and characterised as standardised and clearly structured. Accordingly, we provided different examples to clarify the kind of task in the low-complexity setting.

The complete wording can be seen in Table 3.2.

High task complexity ($n = 705$)	**Low task complexity** ($n = 730$)
Please imagine the following situation: Your company requires additional capacity in the long term for tasks that are typically performed by entry-level employees but require **professional judgement, contextual knowledge, and independent problem-solving**. Examples include interpreting information, preparing individualised recommendations, or handling more complex customer enquiries.	Please imagine the following situation: Your company requires additional capacity in the long term for **standardised, clearly structured tasks** that are typically performed by entry-level employees. Examples include data maintenance, simple research, standardised analyses, or handling recurring customer enquiries.

Table 3.2: Experimental task-complexity manipulation
Notes: Respondents were randomly assigned to one of the two scenarios. The questionnaire versions differed only in the task description shown in the table; the subsequent outcome questions and response scales were identical across treatment groups. The table presents English translations of the original German questionnaire.

The random assignment to one of the two groups took place at the firm level before the questionnaires were printed and distributed. This timing was necessary because the experimental scenario formed part of the printed questionnaire and therefore had to be assigned before questionnaires were sent out. The sample that was contacted in this wave of the survey consists of 11,748 firms, of which only a subset participates in any given wave. A simple random allocation across the full gross sample could therefore

still produce uneven treatment groups among the final respondents. To reduce this risk, we implemented the randomisation in two stages using information about the firms' previous response behaviour.

First, firms that had participated in at least one of the previous two survey waves were randomly assigned to the two questionnaire versions, as those firms have a relatively high probability of participating again. In a second step, the remaining firms in the gross sample were randomised. Before the questionnaires were assigned and sent out, the distribution between both groups regarding firm size and industry were checked.

The procedure can therefore be understood as conditioned random assignment based on recent survey participation. Among the firms participating in the current wave, 705 respondents received the high-complexity questionnaire and 730 received the low-complexity questionnaire.

Table 3.3 presents the balance between both groups on all relevant characteristics. Overall, the groups are well balanced. Among the individual comparisons, only the share of firms with at least 100 employees differs significantly at the 5% level: Within the high complexity group, 32.9% of firms have at least 100 employees, whereas this share amounts to 28.1% in the low-complexity group, a difference of 4.8 percentage points ($p = 0.047$). Given the number of comparisons, one marginally significant difference may arise by chance even under successful random assignment.

More importantly, a joint Wald test does not reject equality of the observed characteristics across treatment groups ($\chi^2(12) = 9.54$, $p = 0.656$). The isolated firm-size difference therefore provides no indication of systematic imbalance. Industry and firm-size fixed effects are added in the adjusted specifications. The treatment estimates remain very similar after these controls are included, showing that the main results are not driven by the somewhat larger share of large firms in the high-complexity group.

3.3 Outcome Variables and Covariates

The two main outcomes were elicited immediately after the experimental scenario and referred explicitly to the type of task described in the respective treatment condition.

3.3.1 AI substitution

The first post-treatment outcome measures the firm's stated likelihood of using AI instead of hiring an additional entry-level employee. The exact wording was:

Variable	Mean High	Mean Low	Diff.	t-stat.	p-value
Information economy	0.53	0.54	-0.01	-0.56	0.5743
Number of employees	207.64	139.24	68.40	1.44	0.1497
5–19 employees	0.30	0.33	-0.03	-1.15	0.2500
20–99 employees	0.37	0.39	-0.02	-0.78	0.4339
100+ employees	0.33	0.28	0.05	1.99	0.0473
Online response	0.73	0.74	-0.01	-0.32	0.7489
Female respondents	0.25	0.25	0.00	-0.10	0.9185
Respondents aged 50+	0.71	0.72	-0.01	-0.29	0.7724
AI applications used	2.24	2.17	0.06	0.80	0.4233
Frequent personal AI use	0.68	0.68	0.00	0.09	0.9285
Any entry-level hires: vocational	0.56	0.53	0.03	1.04	0.2983
Any entry-level hires: university	0.50	0.45	0.05	1.85	0.0644
Hiring difficulty: vocational	3.02	3.02	0.00	0.06	0.9554
Hiring difficulty: university	2.86	2.91	-0.05	-0.62	0.5369

Table 3.3: Balance between the high- and low-task-complexity treatment groups

Notes: The table reports treatment-group means, their difference (high minus low task complexity), and two-sided tests of equality. A joint Wald test of the pre-treatment covariates does not reject joint equality between the treatment groups ($p = 0.656$).

> *How likely would your company be to use AI applications for such tasks instead of hiring an additional entry-level employee?*

Responses were recorded on a five-point scale ranging from 1, "very unlikely", to 5, "very likely". Respondents could alternatively select the outside option "we already do this". Hence, higher scale values indicate a greater willingness to substitute additional entry-level hire by AI.

The main analysis uses the five-point response whenever a scale category is available. This includes the small number of questionnaires on which both a scale category and the outside option were selected. A valid scale response is available for 1,371 firms.

3.3.2 Human oversight

The second main outcome measures the importance managers assign to human oversight of potential AI-generated output:

> *How important would it be to you that AI-generated results for such tasks are critically reviewed by a human?*

The five-point scale ranges from 1, "not at all important", to 5, "very important". The corresponding

outside option was "we generally do not use AI". Higher values indicate greater perceived importance of human oversight. We use all available responses on the five-point scale. This includes the small number of respondents who selected both a scale value and the outside option. Overall, 1,359 firms provided a valid scale response.

3.3.3 Expected hiring effects

Before the experimentally varied scenario, respondents were asked how they expect the availability of AI applications to affect their firm's hiring behaviour over the following two years. The question asked separately for the expected change in hiring behaviour referring to entry-level employees and to employees with professional experience, defined in the questionnaire as workers with established knowledge in their field.

Both outcomes are measured on a fully verbalised five-point scale: 1, "substantially decrease"; 2, "slightly decrease"; 3, "no effect"; 4, "slightly increase"; and 5, "substantially increase". Additionally, respondents could select the outside option "generally no need" for the respective employee group. Because this outside category has no natural position on the directional scale, the tests of Hypotheses 3 and 4 use only responses between 1 and 5. Valid scale responses are available for 1,247 firms for entry-level hiring and 1,291 firms for experienced hiring; 1,230 firms provided valid responses for both questions.

3.3.4 Covariates

The adjusted specifications use variables on firm and respondent characteristics. For firm characteristics, four industry categories and three firm-size classes are used (as listed in Table 3.1). Respondent characteristics comprise indicators for a female respondent and for a respondent aged 50 or older.

Moreover, prior AI experience is measured, both on a firm as well as on a personal level. Firm-level experience is the number of AI application categories[7] currently used by the firm, ranging from zero to six. Personal AI use was elicited by a question regarding the frequency the respondent personally uses AI applications. Frequent personal AI use equals one if the respondent reports using AI applications daily or several times per week.

Subsequently, respondents were asked about their hiring behaviour and difficulties, separately for entry-level employees with a vocational qualification, and for entry-level employees with a university

[7] See Appendix A.1, Q1, for the complete list of response options.

degree. After eliciting the number of respective hires over the past two years, they were asked to assess the difficulty in finding and hiring entry-level employees of both kinds, with the outside option of stating that they generally don't need the respective kind of entry-level employee. The scale ranges from 1, "no difficulty", to 5, "great difficulty". In the regression models, recruitment difficulty is set to zero for firms reporting no need, while separate need indicators distinguish these firms from those reporting a need and low recruitment difficulty.

Two binary indicators capture whether the firm hired at least one entry-level employee with a vocational qualification or a university degree during the previous two years.

The preregistration specified the perceived importance of entry-level employees for the firm's talent pipeline as a pre-treatment measure and a control variable. We subsequently moved this question to after the experimental scenario and both experimental outcomes. This change was motivated by concern that asking about the importance of firm-internal talent development beforehand could prime respondents to consider the long-term consequences of future entry-level hires, potentially altering their response to the task-complexity manipulation. Because the measure was consequently elicited after treatment and could itself be affected by the experimentally varied scenario, we do not include it as a control variable. This represents a deliberate deviation from the preregistration.

Appendix Table B.6 reports the number of observations and treatment allocation in each analytical sample.

4 Empirical Strategy

4.1 Main Specifications

As we use experimental data and randomisation worked well, we estimate the following OLS specification[8] for all experimental outcomes:

$$Y_i = \beta_0 + \beta_1 HighComplexity_i + \beta_2' X_i + \varepsilon_i. \tag{1}$$

where Y_i denotes either the five-point AI-substitution score or the five-point human-oversight score for firm i. The constant β_0 represents the expected outcome in the low-complexity condition when the

[8] All OLS models use heteroskedasticity-robust standard errors (HC3)

included covariates take the value zero (or equal their respective reference category). The indicator $HighComplexity_i$ equals one if firm i received the high-complexity scenario and zero if it received the low-complexity scenario. β_1 is the coefficient of primary interest and captures the average difference in the outcome between the high- and low-complexity conditions. The vector X_i contains the covariates included in the respective specification, while β_2' denotes the corresponding vector of coefficients. Finally, ε_i is the error term.

Hypothesis 1 predicts $\beta_1 < 0$ for AI substitution, whereas Hypothesis 2 predicts $\beta_1 > 0$ for human oversight.

We report four nested specifications. Model (1) is the baseline model and contains only the treatment indicator as an independent variable. Model (2) adds industry and firm-size fixed effects, whereas Model (3) also includes respondent characteristics (gender and age). Model (4) adds firm and personal AI experience, previous entry-level hiring, and the qualification-specific need and recruitment-difficulty measures described in Section 3.3. For all plots and models of our main model specifications we consistently use the set of complete cases used in Model (4). We thereby make sure that changes across columns can be attributed to covariate adjustment and do not stem from a shift in sample composition. The corresponding common samples contain 1,210 observations for AI substitution and 1,205 observations for the model for human oversight.

The hiring-expectation outcomes are not generated by the experimental manipulation. To evaluate Hypotheses 3 and 4, we test the mean response on the 5-point Likert scale against the neutral scale value of 3. This is done for both types of employees separately. We then compare the two expectations within the 1,230 firms that answered both items. This paired comparison holds the respondent constant and tests whether the anticipated change differs between entry-level and experienced hiring. These estimates describe managers' expectations, which are causal in their wording, but not causal measurements on our side.

5 Results

5.1 AI Substitution of Entry-Level Work

Table 5.1 shows the results of the aforementioned OLS model with the managers' assessed probability of using AI applications for the given task instead of hiring an entry-level employee. The baseline model in column (1) of Table 5.1 shows a negative estimation for the main coefficient β_1 that is significant on all conventional levels. Adding firm and respondent characteristics, as well as the pre-specified covariates does not change direction, size or significance of the coefficient estimation. This means that managers are significantly less willing to use AI applications instead of an additional entry-level employee when the task that needs to be done is complex. Thus, the results support Hypothesis 1. If all control variables are accounted for, higher task complexity lowers the AI-substitution score by 0.54 points on the 5-point Likert scale. Figure 5.1 plots the four estimates and their 95% confidence intervals. Further, Model (4) provides additional insights: Higher levels of AI use – both on the firm and on the personal level – correspond with higher likelihoods of using AI instead of hiring an additional entry-level employee. Neither coefficient regarding past difficulties in recruitment is statistically significant at the 5% level. The coefficients on recruitment difficulty and AI use describe conditional associations rather than causal effects.

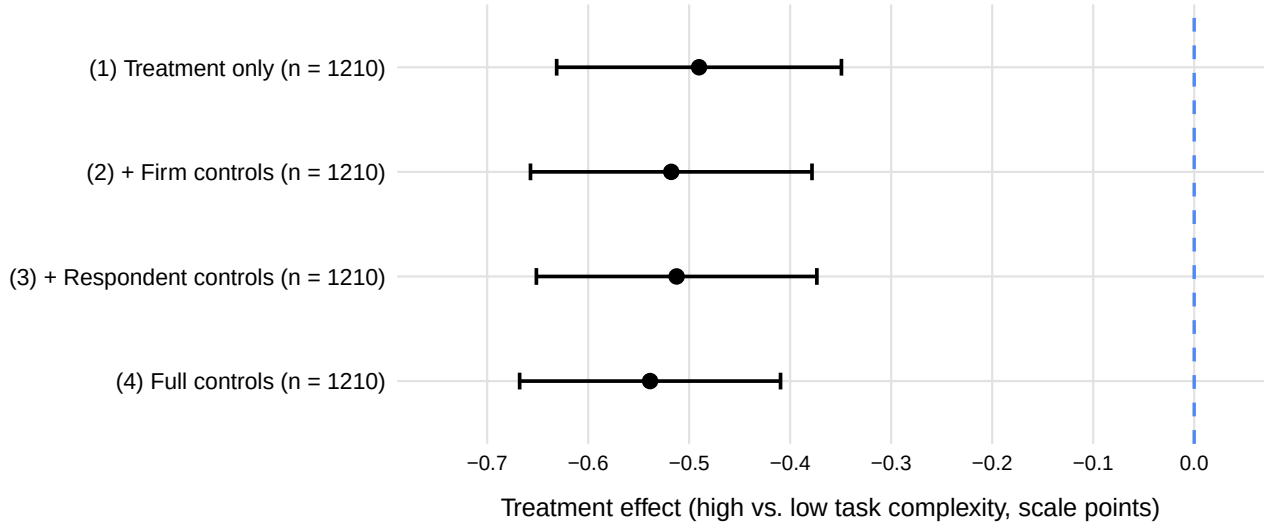


Figure 5.1: Effects of task complexity on AI substitution

Notes: Points show OLS estimates of the effect of high relative to low task complexity on firms' willingness to use AI instead of hiring an entry-level employee. Horizontal lines indicate 95% confidence intervals based on HC3-robust standard errors. All specifications are estimated using the common sample of the full model ($N = 1{,}210$).

Variable	(1)	(2)	(3)	(4)
High task complexity	-0.490***	-0.518***	-0.512***	-0.539***
	(0.072)	(0.071)	(0.071)	(0.066)
AI applications used				0.258***
				(0.028)
Frequent personal AI use				0.389***
				(0.078)
Any previous entry-level hires: vocational				-0.054
				(0.084)
Any previous entry-level hires: university				0.104
				(0.082)
Hiring difficulty: vocational				0.066+
				(0.038)
Hiring difficulty: university				-0.052
				(0.040)
R^2	0.037	0.076	0.087	0.222
Mean dep. var. (low complexity)	3.208	3.208	3.208	3.208
Industry FE	No	Yes	Yes	Yes
Firm size FE	No	Yes	Yes	Yes
Respondent controls	No	No	Yes	Yes
Need indicators	No	No	No	Yes
Observations	1210	1210	1210	1210

Table 5.1: Effects of task complexity on the likelihood of using AI instead of hiring an entry-level employee. HC3-robust standard errors are reported in parentheses. Significance levels: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.2 Human Oversight

Accordingly, Table 5.2 reports the results of the nested OLS models for the assigned importance of human oversight. As before, column (1) of Table 5.2 shows the coefficient β_1 without further controls. The estimation shows a positive effect of the high complexity condition on the importance that is assigned to human oversight. The effect is significant on the 1 percent level. Analogously to the previous models, adding firm and respondent characteristics as well as control variables regarding AI usage and information on previous hiring hardly changes the size of the coefficient of interest. Hence, when managers are presented with a scenario that involves more complex tasks, they assign a higher importance to critical judgement of AI generated outputs by a human. Our results therefore support Hypothesis 2. Figure 5.2 shows the four estimators as well as their confidence bands graphically and displays that the estimation is robust across all four specifications.

The effect of the high complexity condition on human oversight score is considerably smaller than that on the substitution probability. In part because review is considered important even in the low-complexity condition (mean value for the assigned importance (scale 1 to 5) in the low complexity group: 4.32).

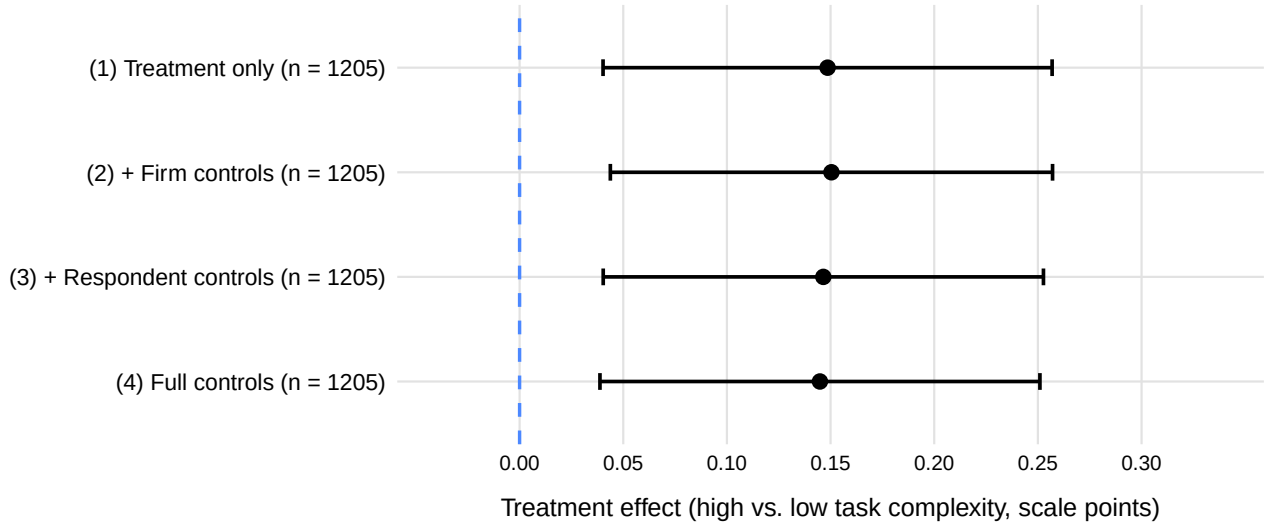


Figure 5.2: Effects of task complexity on the importance of human oversight

Notes: Points show OLS estimates of the effect of high relative to low task complexity on the reported importance of human oversight. Horizontal lines indicate 95% confidence intervals based on HC3-robust standard errors. All specifications are estimated using the common sample of the full model ($N = 1{,}205$).

Variable	(1)	(2)	(3)	(4)
High task complexity	0.149**	0.150**	0.147**	0.145**
	(0.055)	(0.054)	(0.054)	(0.054)
AI applications used				0.064**
				(0.023)
Frequent personal AI use				0.003
				(0.067)
Any previous entry-level hires: vocational				-0.155*
				(0.067)
Any previous entry-level hires: university				0.070
				(0.065)
Hiring difficulty: vocational				-0.029
				(0.029)
Hiring difficulty: university				0.012
				(0.029)
R^2	0.006	0.046	0.056	0.078
Mean dep. var. (low complexity)	4.323	4.323	4.323	4.323
Industry FE	No	Yes	Yes	Yes
Firm size FE	No	Yes	Yes	Yes
Respondent controls	No	No	Yes	Yes
Need indicators	No	No	No	Yes
Observations	1205	1205	1205	1205

Table 5.2: Effects of task complexity on the importance of human oversight. HC3-robust standard errors are reported in parentheses. Significance levels: + $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3 Expected Effects of AI on Hiring

The variables regarding hiring expectation provide a broader view of how managers believe AI availability will affect workforce demand over the next two years. Figure 5.3 reports the response distribution among firms with a stated need for the respective employee group. The majority of firms do not expect a change in hiring behaviour: 63% and 65%, respectively, expect that the availability of AI applications has no effect on their hiring over the next two years.

When looking at the response behaviour of those firms that expect AI availability to change their hiring behaviour, there are substantial differences between entry-level employees and experienced employees. For entry-level workers, 31.3% expect hiring to decrease, whereas 5.5 per cent expect an increase. For experienced employees, the corresponding shares are 24.9% that expect hiring to decrease, whereas 10.1% expect an increase. Thus, "no effect" is the modal response for both groups, but negative expectations are more common for entry-level hiring.

The mean expected effect for entry-level employees is 2.684 (95% CI [2.645, 2.723]), which is significantly below the neutral value of 3 ($t = -15.72$, $p < 0.001$, $N = 1{,}247$). This supports Hypothesis 3 that AI is generally seen as a substitute for entry-level work and that its availability might decrease hiring of these employees.[9]

For experienced employees we expected AI to be a complement, according to Hypothesis 4. Hence, we would expect the mean value for those employees to lie above the neutral value of three. However, the mean for experienced employees is also significantly below 3, at 2.826 (95% CI [2.788, 2.865]; $t = -8.76$, $p < 0.001$, $N = 1{,}291$). Hypothesis 4 is therefore not supported.

Finally, we test whether the hiring expectations for entry-level employees and experienced employees differ statistically significantly. Among the 1,230 firms that answered both items, the expected effect for experienced employees exceeds the expected effect for entry-level employees by 0.144 scale points (95% CI [0.107, 0.181], paired $t = 7.56$, $p < 0.001$). Appendix Table B.3 reports the full test results. Hence, managers anticipate a relative shift towards experienced labour, but not as an absolute expansion of experienced hiring.

[9] This conclusion should be interpreted with caution because our measure captures the expected direction and only a broad indication of the magnitude of adjustment, rather than the number of hires affected. A larger share of firms anticipating a decrease therefore does not necessarily imply a decline in aggregate hiring: this effect could be offset if the smaller group anticipating an increase hires substantially more entry-level employees.

When looking at firms' responses according to their sector and size, it becomes apparent that the expectations to decrease future hires are especially pronounced among large firms. As can be seen from Figure 5.4 43% of the large firms in the information economy expect to reduce their hiring behaviour regarding entry-level workers. That the share of firms expecting declining future hires is the largest among large firms with at least 100 employees is true for both industries and both types of employees. Similarly, the relative shift towards experienced workers can be seen across all firm characteristics.

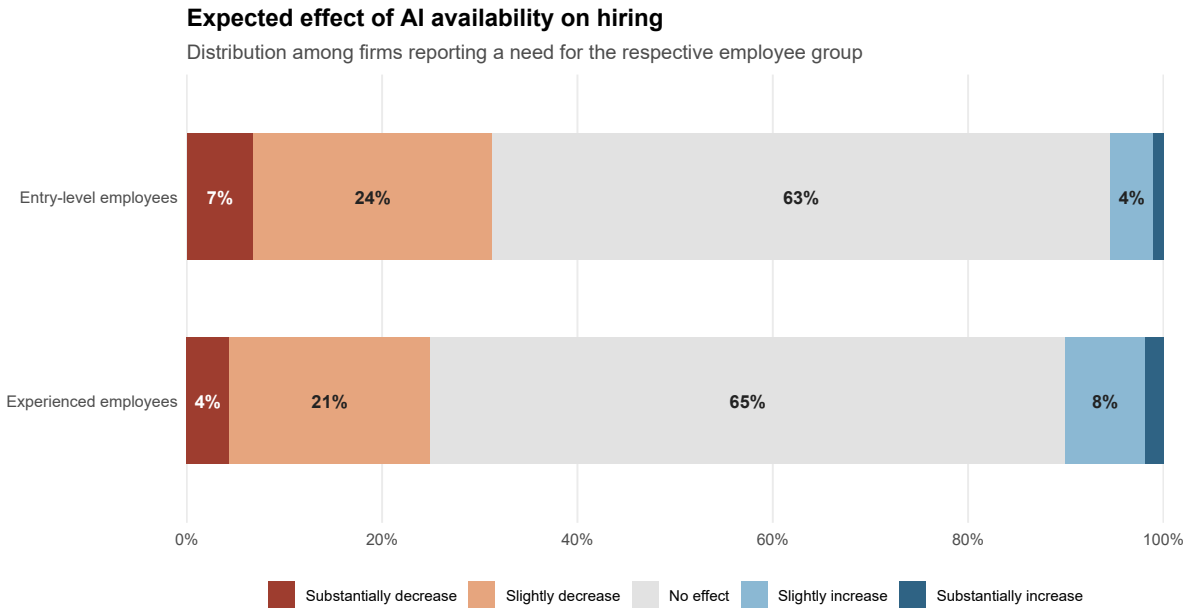


Figure 5.3: Expected effects of AI availability on hiring entry-level and experienced employees

Notes: The figure reports the distribution of managers' expected hiring responses to the availability of AI. The five-point scale ranges from *substantially decrease* to *substantially increase*. Firms reporting no general need for the respective employee group are excluded. Percentages may not sum to 100 due to rounding.

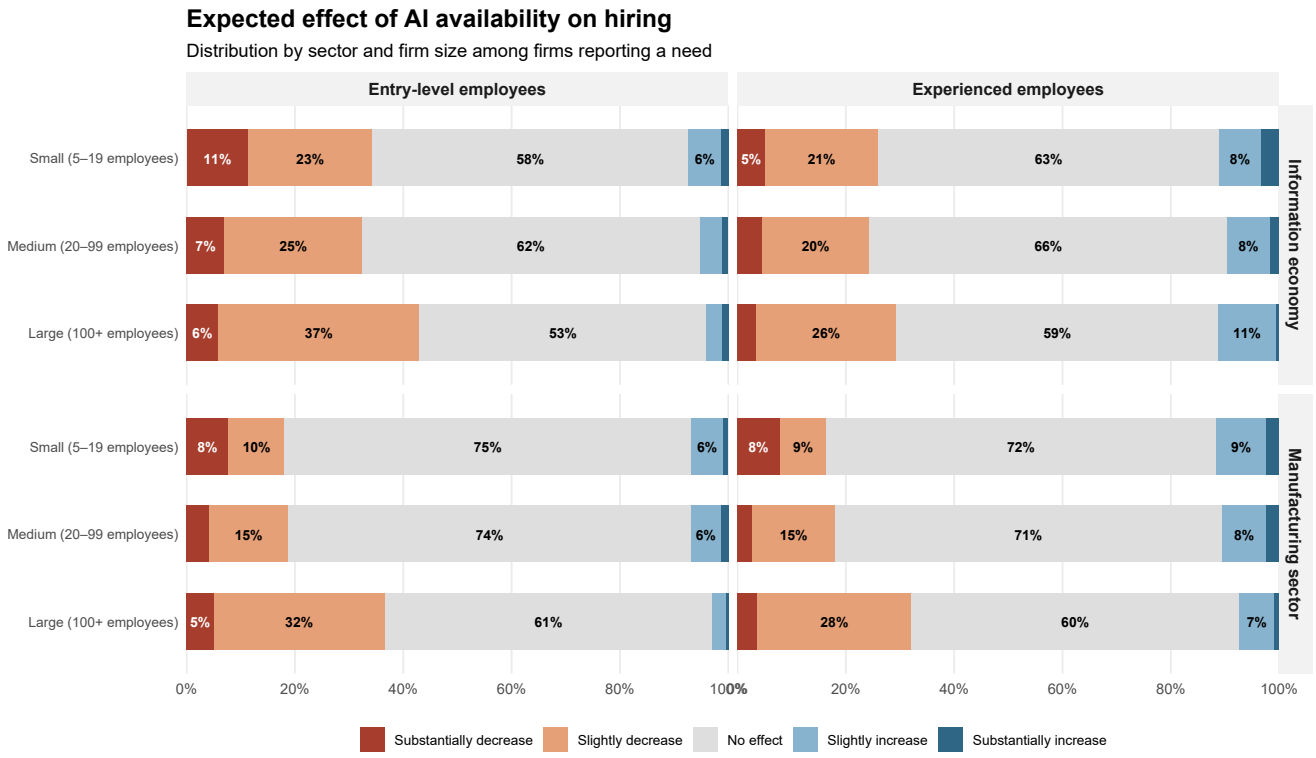


Figure 5.4: Expected effects of AI availability on hiring entry-level and experienced employees, by firm sector and size

Notes: The figure reports the distribution of expected hiring effects separately by sector, firm size, and employee group. Percentages are calculated among firms providing a valid scale response for the respective employee group; firms reporting no general need are excluded. Categories accounting for less than 5% are displayed without a percentage label. Percentages may not sum to 100 due to rounding.

5.4 Hiring Expectations and Business Conditions

Next, we examine whether managers' hiring expectations differ with their firms' recent and anticipated business conditions, i.e. their revenue dynamics. This analysis was not preregistered and should therefore be regarded as exploratory. The ZEW business survey always elicits whether the firms' revenue in the current quarter has risen, stayed equal or has declined compared to the previous quarter. In addition to the revenue development in the current (i.e. past as we send out the survey at the end of a quarter) quarter, they are asked about their revenue expectations for the next quarter. Again, the options are increase, equal, decrease. For our analysis, we distinguish three mutually exclusive groups: Firms are classified as experiencing *sustained revenue growth* if they report both a recent increase and an expected further increase in revenue. *Sustained revenue decline* analogously requires both a recent decrease and an expected further decrease. All remaining firms with valid responses to both revenue questions are classified as having *other revenue dynamics*.

Figure 5.5 presents the complete response distributions for these three groups. As in the previous analyses, the samples are restricted to firms reporting a general need for the respective employee group. The entry-level sample comprises 150 firms experiencing sustained revenue growth, 948 firms with other revenue dynamics, and 129 firms experiencing sustained revenue decline. The corresponding numbers for experienced employees are 154, 978, and 139. Appendix Figure B.1 provides a more detailed decomposition of the middle group.

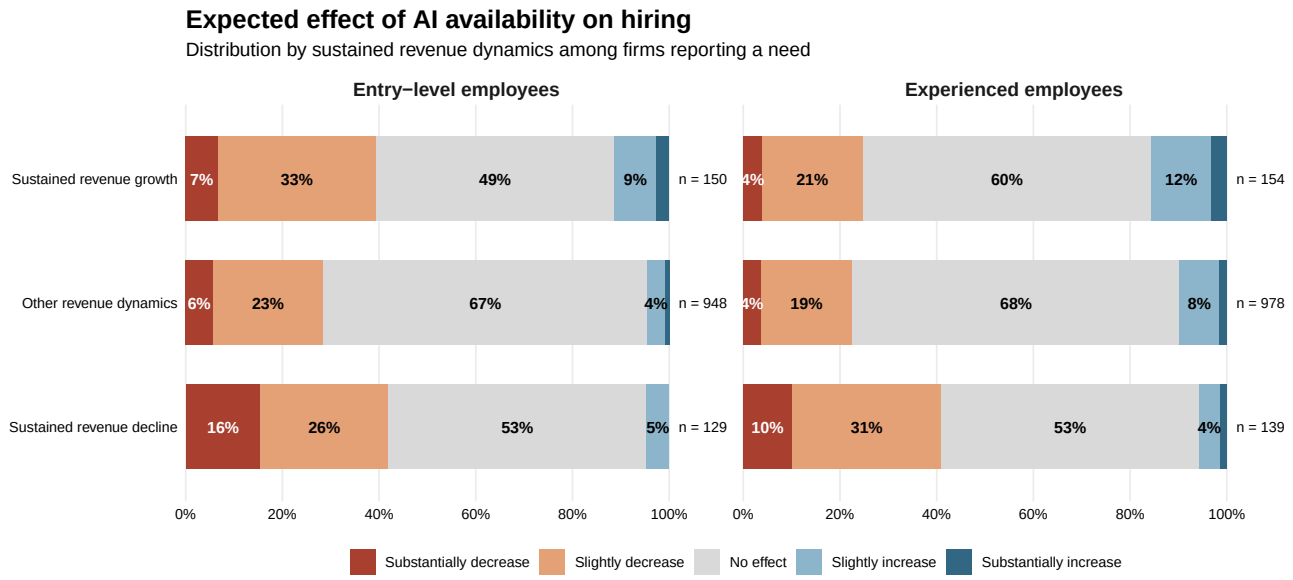


Figure 5.5: Expected effects of AI availability on hiring, by revenue dynamics

Notes: The figure reports managers' expected effects of AI availability on hiring. Sustained revenue growth denotes firms reporting both an increase in revenue during the preceding quarter and an expected increase during the following quarter. Sustained revenue decline analogously requires decreases in both periods. Firms with all remaining valid combinations are classified as exhibiting other revenue dynamics. Firms reporting no general need for the respective employee group are excluded. The number of valid observations is shown next to each bar. Percentages may not sum to 100 due to rounding.

Firms experiencing sustained revenue decline report substantially more negative hiring expectations. For entry-level employees, 41.9% of these firms expect AI availability to reduce hiring, compared with 28.5% of firms exhibiting other revenue dynamics. The difference is also visible at the lower endpoint of the scale: 15.5% of firms experiencing sustained decline expect entry-level hiring to decrease substantially, compared with 5.6% of the middle group. For experienced employees, the share expecting any decrease is 41.0% among firms experiencing sustained revenue decline and 22.6% among firms with other revenue dynamics. The respective shares expecting a substantial decrease are 10.1% and 3.6%. Hence, the share of firms expecting substantial decreases in hiring entry-level employees due to the availability of AI applications more than doubles among firms that had their revenue declining in the past quarter and expect a further decline in the next one.

Comparisons based on the complete five-point scale confirm this pattern. The mean entry-level response is 2.473 among firms experiencing sustained revenue decline and 2.713 among firms with other revenue dynamics, yielding a difference of -0.240 scale points (HC3 SE $= 0.075$, 95% CI $[-0.387, -0.094]$, $p = 0.001$). For experienced employees, the respective means are 2.561 and 2.852, corresponding to a

difference of −0.291 points (HC3 SE = 0.071, 95% CI [−0.429, −0.152], $p < 0.001$). Both differences remain statistically significant after jointly adjusting the four unadjusted group comparisons using Holm's sequentially rejective procedure (Holm, 1979) ($p_{\text{Holm}} = 0.004$ and $p_{\text{Holm}} < 0.001$, respectively).

These differences are not driven by the sector or size composition of the groups. After including industry and firm-size fixed effects, sustained revenue decline is associated with a reduction of 0.275 scale points in the entry-level outcome (HC3 SE = 0.074, $p < 0.001$) and 0.304 points in the experienced-employee outcome (HC3 SE = 0.072, $p < 0.001$). Both adjusted estimates also remain significant after Holm correction. Appendix Table B.4 reports the unadjusted and adjusted mean comparisons for both employee groups.

The pattern for firms experiencing sustained revenue growth is more nuanced. These firms are descriptively more likely to expect increases in hiring due to AI availability. For entry-level employees, 11.3% anticipate an increase, compared with 4.5% among firms with other revenue dynamics. For experienced employees, the corresponding shares are 15.6% and 9.7%. At the same time, particularly for entry-level hiring, sustained-growth firms are also more likely to expect a decrease. Consequently, only 49% of sustained-growth firms select "no effect" for entry-level employees, compared with 67% in the middle group. The distribution therefore becomes more polarised in both directions rather than shifting uniformly towards more favourable hiring expectations.

Consistent with this interpretation, sustained revenue growth is not associated with a statistically significant difference in the mean response. The estimated difference relative to firms with other revenue dynamics is −0.033 scale points for entry-level employees ($p = 0.643$) and 0.051 points for experienced employees ($p = 0.447$). With industry and firm-size fixed effects, the corresponding estimates are −0.006 and 0.070, and both remain statistically insignificant.

Binary comparisons provide suggestive, but not conclusive, evidence of higher expected hiring increases among firms in great business conditions, i.e. having increasing revenue and expecting further revenue increase for the near future. For entry-level employees, the 6.8-percentage-point difference in the probability of expecting any increase is statistically significant before correcting for multiple testing ($p = 0.012$), but falls just outside the 5% level after applying the Holm correction across the eight reported threshold comparisons ($p_{\text{Holm}} = 0.058$). For experienced employees, the difference amounts to 5.9 percentage points ($p = 0.058$; $p_{\text{Holm}} = 0.173$). There is no statistically significant difference in the

probability of expecting a substantial increase. The pattern for the firms belonging to the growth group should therefore be interpreted as descriptive evidence rather than evidence of an unambiguously positive shift in expected hiring.

The binary comparisons reinforce the sustained-decline result. After Holm correction across the same eight threshold tests, the difference in the probability of expecting any decrease remains significant for both entry-level employees ($p_{\text{Holm}} = 0.023$) and experienced employees ($p_{\text{Holm}} < 0.001$). The difference in the probability of expecting a substantial decrease remains significant for entry-level employees ($p_{\text{Holm}} = 0.019$), while the corresponding comparison for experienced employees falls just outside the 5% threshold ($p_{\text{Holm}} = 0.058$). Appendix Table B.5 reports all eight response-share comparisons, including the unadjusted and Holm-adjusted p-values.

Overall, the results indicate a robust association between sustained revenue decline and more negative perceived employment consequences of AI. The evidence for sustained-growth firms instead suggests that managers in economically expanding firms are more likely to anticipate some adjustment in hiring, but disagree about its direction. These relationships should not be interpreted causally. Firms' revenue dynamics were not randomly assigned, and current business conditions may capture broader differences in demand, investment, restructuring, or managerial expectations.

6 Additional Analyses and Robustness

6.1 Treatment Effect Heterogeneity

This section provides heterogeneity analyses. We examine whether the treatment effects vary with firm size, sector, firm-level AI use, frequent personal AI use, previous entry-level hiring, and recruitment difficulty. Appendix Figures B.2 and B.3 report subgroup-specific effects estimated on the respective full-control sample, together with 95% confidence intervals. Formal interaction tests, rather than differences in subgroup significance levels, are used to assess heterogeneity.

For AI substitution, the estimated treatment effect is negative in every displayed subgroup. The joint interaction tests do not reject equality by firm size ($p = 0.430$), sector ($p = 0.304$), firm AI use ($p = 0.361$), personal AI use ($p = 0.187$), previous entry-level hiring ($p = 0.103$), or recruitment difficulty ($p = 0.631$). The estimated treatment effects on human oversight are less precise across subgroups, as

expected for the smaller average effect, but again yield no statistically significant interaction.

The data therefore provide no evidence that the response to task complexity is restricted to a particular type of firm or manager. This should not be read as proof of homogeneous effects. Our conclusion is that the average treatment effects remain visible across the observed sample, while systematic moderation is not detected by any of our explanatory variables.

6.2 Robustness Checks

Our main specifications treat the five-point outcome scales as numerical variables. We additionally estimate ordered probit models to account for their ordinal nature. As shown in Appendix Table B.1, the fully adjusted estimates are negative for AI substitution and positive for human oversight, with both treatment coefficients statistically significant at the 0.1 per cent level. The models use the same controls and the corresponding common samples as the fully adjusted OLS specifications. Inference is based on heteroskedasticity-robust sandwich standard errors. Ordered probit coefficients refer to a latent index and cannot be compared directly with the OLS coefficients in scale points, but the direction and statistical significance of the treatment effects are consistent.

Moreover, our results remain consistent when we exclude ambiguous responses, i.e. responses that combined scale values with the outside option. For instance, when a firm chose a value between 1 and 5 on how important they see human oversight of AI output, and simultaneously chose the outside option that they do generally not use AI, their provided scale value was used in our analyses. All results remain valid if these ambiguous responses are excluded. Further, our reported results used common sample specifications, i.e. all columns of Table 5.1 and 5.2 use the same sample, that is restricted by the most complex model in column (4) of both tables. Baseline results are not driven by this sample selection and remain stable in their size and significance if the full available sample is used for each model separately.

7 Discussion

7.1 Interpretation and Contributions

Our experiment isolates a feature of managerial decision-making that observational measures of AI exposure cannot identify. The results identify task content as an important determinant of managers'

willingness to use AI instead of hiring an entry-level employee. Both experimental scenarios describe a lasting need for additional capacity and refer to work typically performed by entry-level employees. What changes is the description of the tasks. Managers are less willing to substitute AI for an additional hire when the work requires professional judgement, contextual knowledge, and independent problem-solving. The experiment therefore shows that entry-level status alone is not sufficient to characterise the perceived substitutability of work.

The human-oversight result complements this finding. Managers attach greater importance to critical review of AI output when tasks are complex, although oversight is rated as highly important in both conditions. This suggests that task complexity matters not only for whether managers consider AI an alternative to hiring, but also for the human involvement they consider necessary when AI is used.

Our pre-treatment items on hiring expectations for the next two years provide further insights on how AI might shape the future labour market. Managers expect AI availability to reduce hiring of both entry-level and experienced employees, with more negative expectations for entry-level workers. The evidence therefore supports a relative difference between career stages, rather than an absolute increase in demand for experienced employees. These responses concern anticipated hiring decisions and should not be interpreted as a forecast of aggregate employment losses.

The exploratory analysis of revenue dynamics adds a further qualification. Firms reporting a recent revenue decline and expecting a further decline have more negative AI-related hiring expectations for both employee groups. Firms reporting sustained growth are more likely to expect hiring increases, but, particularly for entry-level employees, are also more likely to expect decreases compared to firms that do not belong to the two extreme groups of sustained revenue growth or sustained revenue decline. Favourable business conditions therefore do not translate into a clearly more positive mean hiring expectation. These associations suggest that expectations about AI are formed in the context of firms' broader economic circumstances.

For firms and policymakers, a central question is whether the automation of entry-level tasks changes opportunities to acquire experience. If standardised tasks provide an initial route into an occupation, using AI instead of recruiting juniors could make that route narrower. Training and work organisation may then need to provide other ways for new entrants to develop the judgement and contextual knowledge that managers continue to value.

7.2 Limitations

This study has several limitations. Most importantly, our outcomes are stated managerial choices in a hypothetical scenario. The random assignment identifies how the task description changes those choices, but it does not show how these firms would respond in reality. Hiring decisions involve firm strategy, wages, implementation costs, applicant quality, regulation and many more variables that we cannot control in our setting. Secondly, respondents were asked about "AI applications" rather than a single named technology. This wording allows for different interpretations among respondents and managers may have had different systems and capabilities in mind. Yet, our random assignment protects the treatment comparison if those interpretations are balanced across conditions.

Third, external validity is bounded by the sampling frame. The data cover German firms with at least five employees in the information economy and manufacturing sector, and many respondents have already had substantial exposure to AI prior to the survey. Our findings might be different for micro-firms, which constitute a large part of the German economy. Similarly, they do not automatically extend to other industries, countries or labour markets. Moreover, AI capabilities and organisational practices are changing rapidly. The estimates describe managerial judgements at the time of the 2026 survey and should be treated as such.

Finally, the hiring-expectation results are forecasts rather than realised employment changes. Common shocks to demand or respondents' general optimism may influence both groups of employees, and the comparison cannot establish why managers expect fewer hires. Its strength lies in the within-firm contrast and in the distinction between career stages. Causal evidence on actual hiring will require longitudinal firm data or experiments tied more closely to real vacancies.

8 Conclusion

This paper uses a preregistered survey experiment with managers at 1,435 German firms to investigate how task complexity affects their willingness to use AI instead of hiring an additional entry-level employee. Our setting allows us to experimentally vary task complexity in a hypothetical hiring scenario and elicit the likelihood of managers using AI applications instead of hiring additional entry-level employees as well as the importance they assign to human oversight of AI output. Our results support the pre-

registered hypotheses that a higher level of task complexity leads to a lower likelihood of substituting entry-level labour with AI while it also increases the desire for critical human oversight. Both treatment coefficients have the expected sign and are statistically significant at the 1% level in the common-sample OLS specifications. They remain statistically significant at least at the 5% level across several robustness checks and are also supported by the ordered probit estimates.

Our results support the task-based view of technological change discussed earlier in the paper. Managers distinguish between different kinds of entry-level work, even when the career stage of the employee and the need for additional capacity are held constant. Standardised tasks are more readily considered suitable for substitution by AI, whereas complex tasks retain a stronger role for human judgement and review.

In addition to the two pre-registered hypotheses regarding the effect of the high-complexity setting on these post-treatment variables, we hypothesised that AI might be seen as a substitute to entry-level labour, but as a complement to experienced labour. While our data confirm the hypothesis that AI acts as a substitute for junior work, we have to reject the hypothesis that AI and experienced employees act as complements, i.e. that the availability of AI tools may lead to an increase in firms' hiring expectations regarding more experienced employees. Instead, AI availability also decreases hiring expectations for experienced workers over the next two years. However, the expected reduction for them is significantly lower compared to the entry-level workers. Hence, we still see a relative shift in hirings towards experienced employees. We do not observe real labour market outcomes, but only managerial expectations for firms' future hiring behaviour. Exploratory analyses provide no statistically significant evidence that the treatment effect varies with firms' or managers' prior AI use, previous recruitment difficulties, or previous entry-level hiring. Higher firm-level and personal AI use are nevertheless positively associated with stated willingness to substitute AI for an additional entry-level hire.

Expectations to reduce hiring are especially pronounced among large firms and those who face economic challenges, i.e. suffer declining revenue and expect their revenue to decline further. Firms reporting and expecting revenue growth give more mixed responses, with no statistically significant difference in mean hiring expectations compared with the remaining revenue group, while descriptively exhibiting higher expectations of increasing hiring behaviour compared to firms with less prosperous revenue settings.

The overall sentiment of managers that the availability of AI tools will lead to a decrease in hiring – not only but especially – entry-level employees might have far-reaching consequences regarding both the labour market and the firms' own talent pipeline. When the willingness to hire entry-level workers declines, applicants without major expertise will have it much harder to enter the labour market. Additional trainings and labour market entry mechanisms might become necessary. On the other side, when firms require expertise in order to critically review AI-generated outputs, but refrain from hiring junior level workers, there might be a lack of experienced workers in the long run. Many firms rely on internal talent pipelines, growing their own experienced workforce. If this mechanism is disrupted, potentially for cost savings in the short run, firms may face difficulties developing the experienced workforce they need in the longer run. Managers as well as policy-makers should be aware of this conflict and the potential consequences firms might face in the future. A challenge for both is to realise and use the benefits of AI while maintaining routes through which new entrants can acquire the judgement and experience that complex work continues to require.

References

Acemoglu, D. and P. Restrepo (2019). "Automation and new tasks: How technology displaces and reinstates labor". In: *Journal of Economic Perspectives* 33.2, pp. 3–30.

Autor, D. H. (2013). "The "task approach" to labor markets: An overview". In: *Journal for Labour Market Research* 46.3, pp. 185–199.

Brynjolfsson, E., B. Chandar, and R. Chen (Aug. 2026). *Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence*. Working Paper. Revised version; original version published in 2025. Stanford Digital Economy Lab. URL: https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/.

Brynjolfsson, E., D. Li, and L. Raymond (May 2025). "Generative AI at work". In: *The Quarterly Journal of Economics* 140.2, pp. 889–942.

Carpanelli, M., J. Duszynski, and F. Stephany (Aug. 2026). *Navigating the skill diversity frontier: How skill complexity explains worker resilience*. Preprint arXiv:2608.02102. arXiv. URL: https://arxiv.org/abs/2608.02102.

Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Krayer, F. Candelon, and K. R. Lakhani (Sept. 2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality*. Working Paper 24-013. Harvard Business School. URL: https://aiinstitute.hbs.edu/navigating-the-jagged-technological-frontier/.

Fairlie, R. W. and J. Wu (Sept. 2026). *The early impacts of AI on employment among recent college graduates*. NBER Working Paper 35796. National Bureau of Economic Research.

Hampole, M., D. Papanikolaou, L. D. W. Schmidt, and B. Seegmiller (Feb. 2025). *Artificial intelligence and the labor market*. NBER Working Paper 33509. Revised September 2025. National Bureau of Economic Research.

Holm, S. (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70. URL: https://www.jstor.org/stable/4615733.

Lambert, P. J. and Y. Schindler (May 2026). *The broken ladder: AI, remote work, and early-career hiring*. Warwick Economics Research Paper 1615. University of Warwick, Department of Economics. URL: https://wrap.warwick.ac.uk/202349/.

Lodefalk, M., L. Löthman, M. Koch, and E. Engberg (2026). *Same storm, different boats: Generative AI and the age gradient in hiring*. Ratio Working Paper 388. Ratio Institute. URL: https://ratio.se/publikationer/same-storm-different-boats-generative-ai-and-the-age-gradient-in-hiring.

Mäkelä, E., M. Bone, M. Sehrer, F. Nanji, and F. Stephany (June 2026). *Complement or substitute? How AI increases the demand for human skills*. Preprint arXiv:2412.19754. Version 4, revised June 2026; first posted in 2024. arXiv. URL: https://arxiv.org/abs/2412.19754.

Noy, S. and W. Zhang (2023). "Experimental evidence on the productivity effects of generative artificial intelligence". In: *Science* 381.6654, pp. 187–192.

OECD (2026). *OECD Employment Outlook 2026: Geographic disparities in jobs and incomes*. Paris: OECD Publishing.

Tucker, L. C. (Apr. 2026). *You're (not) hired: Artificial intelligence and early career hiring in the Quarterly Workforce Indicators*. CES Working Paper CES-26-27. U.S. Census Bureau, Center for Economic Studies. URL: https://www.census.gov/library/working-papers/2026/adrm/CES-WP-26-27.html.

Westby, S., A. S. Modestino, and P. Cheng (June 2026). *Generative AI and the redefinition of entry-level software work*. IZA Discussion Paper 18723. IZA Institute of Labor Economics. URL: https://www.iza.org/publications/dp/18723/generative-ai-and-the-redefinition-of-entry-level-software-work.

Appendix

A Survey Instrument and Experimental Materials

This appendix provides an English translation of the survey questions used in the present study. The original questionnaire was administered in German. Questions unrelated to the present study, including the regular business-cycle and working-from-home modules of the ZEW survey, are not reproduced here.

A.1 Firm-Level and Personal AI Experience

Q1. Which of the following artificial intelligence (AI) applications are used in your company?

For each application, respondents could select *Yes* or *No*.

- Analysis or processing of text, speech, and image material
- Generative AI, for example for producing text or computer code
- Machine learning for data analysis and forecasting
- AI-supported workflow automation, for example invoice processing or customer-service chatbots
- Autonomous physical systems, for example warehouse and production robots or driverless transport systems
- Other AI applications

Q2. Approximately how often do you personally use AI applications?

- Daily
- Several times per week
- Several times per month
- Approximately once per month
- Less than once per month
- Never

A.2 Previous Entry-Level Hiring and Hiring Difficulties

Q3. Approximately how many entry-level employees has your company hired over the past two years?

Entry-level employees are defined as individuals taking up employment in their field for the first time, for example after completing vocational training or a university degree. This also includes retaining the company's own apprentices as employees. Respondents were asked to enter zero if no entry-level employees had been hired.

- Number of entry-level employees with a vocational qualification: _____
- Number of entry-level employees with a university degree: _____

Q4. Does your company generally experience difficulties finding and hiring entry-level employees?

Respondents answered separately for the following qualification groups:

- Entry-level employees with a vocational qualification
- Entry-level employees with a university degree

Response scale: 1 = no difficulties; 2; 3; 4; 5 = major difficulties.

Outside option: Generally no need.

A.3 Expected Effects of AI on Hiring

Q5. How do you think the availability of AI applications will affect the number of employees hired by your company over the next two years?

Respondents answered separately for:

- Entry-level employees
- Experienced employees, that is, employees with sound knowledge in their field

Response options:

1. Substantially decrease
2. Slightly decrease
3. No effect
4. Slightly increase
5. Substantially increase

Outside option: Generally no need.

A.4 Experimental Manipulation

Respondents were randomly assigned to one of the following two scenarios.

High Task Complexity

> Please imagine the following situation: Your company requires additional capacity in the long term for tasks that are typically performed by entry-level employees but require professional judgement, contextual knowledge, and independent problem-solving. Examples include interpreting information, preparing individualised recommendations, or handling more complex customer enquiries.

Low Task Complexity

> Please imagine the following situation: Your company requires additional capacity in the long term for standardised, clearly structured tasks that are typically performed by entry-level employees. Examples include data maintenance, simple research, standardised analyses, or handling recurring customer enquiries.

A.5 Primary Experimental Outcomes

Each respondent received the following questions immediately after the randomly assigned scenario.

Q6. How likely would your company be to use AI applications for such tasks instead of hiring an additional entry-level employee?

Response scale: 1 = very unlikely; 2; 3; 4; 5 = very likely.

Outside option: We already do this.

Q7. How important would it be to you that AI-generated results for such tasks be critically reviewed by a human?

Response scale: 1 = not at all important; 2; 3; 4; 5 = very important.

Outside option: We generally do not use AI.

A.6 Perceived Importance of Entry-Level Employees

Q8. How important are entry-level employees to your company for developing future skilled workers and managers?

Response scale: 1 = not at all important; 2; 3; 4; 5 = very important.

A.7 Firm and Respondent Characteristics

Q9. How many employees did your company have in Germany on average in 2025, including management and apprentices?

Number of employees: _____

Share holding a university degree: _____%

Q10. What is your age?

- Under 30
- 30–39
- 40–49
- 50–60
- Over 60

Q11. What is your gender identity?

- Female
- Male
- Other
- Prefer not to answer

B Additional Tables and Figures

B.1 Alternative Estimators and Robustness Checks

Variable	(1)	(2)
Outcome	AI substitution	Human oversight
High task complexity	-0.501***	0.252***
	(0.063)	(0.070)
Industry FE	Yes	Yes
Firm size FE	Yes	Yes
Respondent controls	Yes	Yes
AI experience controls	Yes	Yes
Previous hiring controls	Yes	Yes
Need and difficulty controls	Yes	Yes
Observations	1210	1205

Table B.1: Ordered probit estimates of the effects of task complexity on AI substitution and human oversight

Notes: The table reports ordered probit coefficients for the two five-point primary outcomes. Both specifications include the complete set of controls used in the fully adjusted main models and use the corresponding common estimation samples. Heteroskedasticity-robust sandwich standard errors are reported in parentheses. Coefficients refer to the latent index and are not directly comparable with the OLS coefficients in scale points. Significance levels: $+\ p < 0.10$, $*\ p < 0.05$, $**\ p < 0.01$, $***\ p < 0.001$.

B.2 Response Patterns and Outside Options

Table B.2: Response patterns and outcome availability by treatment assignment

Response pattern	High complexity	Low complexity	Difference (pp)	HC3 SE (pp)	p-value	Observations
Panel A: AI substitution						
Scale only	657 (93.2%)	668 (91.5%)	1.68	1.40	0.230	1435
Outside option only	18 (2.6%)	24 (3.3%)	-0.73	0.89	0.409	1435
Scale and outside option	15 (2.1%)	31 (4.2%)	-2.12	0.92	0.022	1435
No answer	15 (2.1%)	7 (1.0%)	1.17	0.65	0.074	1435
Any outside-option selection	33 (4.8%)	55 (7.6%)	-2.82	1.28	0.027	1413
Any valid scale response	672 (95.3%)	699 (95.8%)	-0.43	1.09	0.691	1435
Panel B: Human oversight						
Scale only	669 (94.9%)	679 (93.0%)	1.88	1.26	0.135	1435
Outside option only	23 (3.3%)	34 (4.7%)	-1.40	1.03	0.175	1435
Scale and outside option	4 (0.6%)	7 (1.0%)	-0.39	0.46	0.394	1435
No answer	9 (1.3%)	10 (1.4%)	-0.09	0.60	0.877	1435
Any outside-option selection	27 (3.9%)	41 (5.7%)	-1.82	1.13	0.110	1416
Any valid scale response	673 (95.5%)	686 (94.0%)	1.49	1.18	0.208	1435

Notes: Entries in the treatment columns report the number and percentage of respondents exhibiting the respective response pattern. Differences are calculated as high task complexity minus low task complexity and are reported in percentage points. HC3 heteroskedasticity-robust standard errors are based on linear probability models. "Any outside-option selection" includes respondents who selected only the outside option or both the outside option and a scale value; respondents providing no answer are excluded from this comparison. "Any valid scale response" includes respondents who provided a scale value, including those who additionally selected an outside option.

B.3 Hiring Expectations

Table B.3: Expected effects of AI availability on hiring entry-level and experienced employees

Outcome or comparison	N	Estimate	SD	95% CI	Reference	t-statistic	p-value
Panel A: Expected hiring effects by employee group							
Entry-level employees	1247	2.684	0.710	[2.645, 2.723]	3	-15.722	< 0.001
Experienced employees	1291	2.826	0.712	[2.788, 2.865]	3	-8.762	< 0.001
Panel B: Paired comparison between employee groups							
Experienced minus entry-level	1230	0.144	0.667	[0.107, 0.181]	0	7.561	< 0.001

Notes: Hiring expectations are measured on a five-point scale ranging from 1 (*substantially decrease*) to 5 (*substantially increase*); 3 indicates no expected effect. Firms reporting no general need for the respective employee group are excluded. Panel A reports two-sided one-sample t-tests against the neutral scale value of 3. Panel B is based on firms providing valid answers for both employee groups and reports the paired difference between experienced and entry-level employees. A positive paired difference indicates relatively more favourable hiring expectations for experienced employees. The reference value for the paired test is zero. In Panel B, SD denotes the standard deviation of the within-firm differences.

B.4 Detailed Revenue Dynamics

Figure B.1 provides a more detailed decomposition of the revenue classification used in Section 5.4. The groups for sustained revenue growth and sustained revenue decline are unchanged. The middle category from the main analysis is divided into firms reporting no decline in either period (i.e. the current/past and next quarter) but not increases in both, and firms reporting a decline in exactly one of the two periods. This four category presentation documents the complete distribution of hiring expectations without changing the focused comparisons reported in the main text.

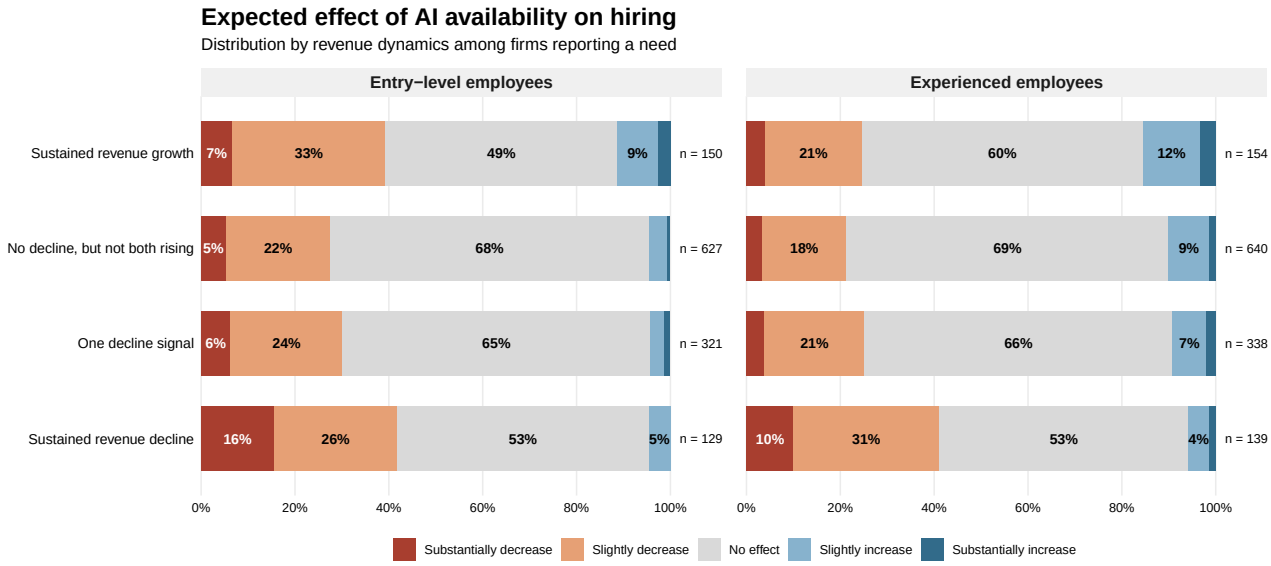


Figure B.1: Expected effects of AI availability on hiring, by detailed revenue dynamics

Notes: The figure reports managers' expected effects of AI availability on hiring across four mutually exclusive combinations of recent and expected revenue development. Sustained revenue growth requires increases in both periods, while sustained revenue decline requires decreases in both periods. Other non-declining revenue dynamics comprise the remaining firms reporting no decrease in either period. One revenue decline signal denotes firms reporting a decrease in exactly one period. Firms reporting no general need for the respective employee group are excluded. The number of valid observations is reported for each bar. Percentages may not sum to 100 due to rounding. The classification and associated analyses were not preregistered and should be interpreted as exploratory.

Specification	Revenue group	N	Difference	HC3 SE	p	p_{Holm}
Panel A: Entry-level employees						
Unadjusted	Sustained growth	1227	-0.033	0.071	0.643	0.893
Unadjusted	Sustained decline	1227	-0.240	0.075	0.001	0.004
Adjusted	Sustained growth	1227	-0.006	0.072	0.934	0.934
Adjusted	Sustained decline	1227	-0.275	0.074	< 0.001	< 0.001
Panel B: Experienced employees						
Unadjusted	Sustained growth	1271	0.051	0.067	0.447	0.893
Unadjusted	Sustained decline	1271	-0.291	0.071	< 0.001	< 0.001
Adjusted	Sustained growth	1271	0.070	0.067	0.297	0.594
Adjusted	Sustained decline	1271	-0.304	0.072	< 0.001	< 0.001

Table B.4: Hiring expectations by revenue dynamics: differences in mean responses

Notes: The table reports differences in mean hiring-expectation scores between each listed revenue group and firms with other revenue dynamics. Sustained growth denotes a reported revenue increase in the second quarter of 2026 relative to the previous quarter and an expected increase in the third quarter 2026; sustained decline analogously requires decreases. The reference group comprises all remaining valid combinations. Hiring expectations range from 1 (substantially decrease) to 5 (substantially increase), with 3 indicating no effect. Firms reporting no general need for the respective employee group are excluded. Adjusted models include industry and firm-size fixed effects. N denotes the total estimation sample, including all three revenue groups. Standard errors are heteroskedasticity-robust (HC3). All tests are two-sided. Holm-adjusted p-values account jointly for four comparisons within each specification: growth and decline relative to the reference group for each employee group. The analysis is exploratory.

Response	Revenue group	Other (%)	Group (%)	Diff. (pp)	SE (pp)	p	p_{Holm}
Panel A: Entry-level employees							
Any decrease	Sustained decline	28.5	41.9	13.4	4.6	0.004	0.023
Substantially decrease	Sustained decline	5.6	15.5	9.9	3.3	0.003	0.019
Any increase	Sustained growth	4.5	11.3	6.8	2.7	0.012	0.058
Substantially increase	Sustained growth	0.8	2.7	1.8	1.4	0.180	0.359
Panel B: Experienced employees							
Any decrease	Sustained decline	22.6	41.0	18.4	4.4	< 0.001	< 0.001
Substantially decrease	Sustained decline	3.6	10.1	6.5	2.6	0.014	0.058
Any increase	Sustained growth	9.7	15.6	5.9	3.1	0.058	0.173
Substantially increase	Sustained growth	1.6	3.2	1.6	1.5	0.281	0.359

Table B.5: Hiring expectations by revenue dynamics: differences in response shares

Notes: Other and Group report response shares among firms with other revenue dynamics and the listed revenue group, respectively. Differences are calculated as the listed group minus the reference group and reported in percentage points (pp). Revenue groups are defined as in Table B.4. Any decrease combines responses 1 and 2; substantially decrease denotes response 1. Any increase combines responses 4 and 5; substantially increase denotes response 5. Firms reporting no general need for the respective employee group are excluded. Differences and HC3-robust standard errors are obtained from unadjusted linear probability models with revenue-group indicators. All tests are two-sided. Holm-adjusted p-values account jointly for all eight comparisons reported in the table. The analysis is exploratory.

B.5 Treatment Effect Heterogeneity

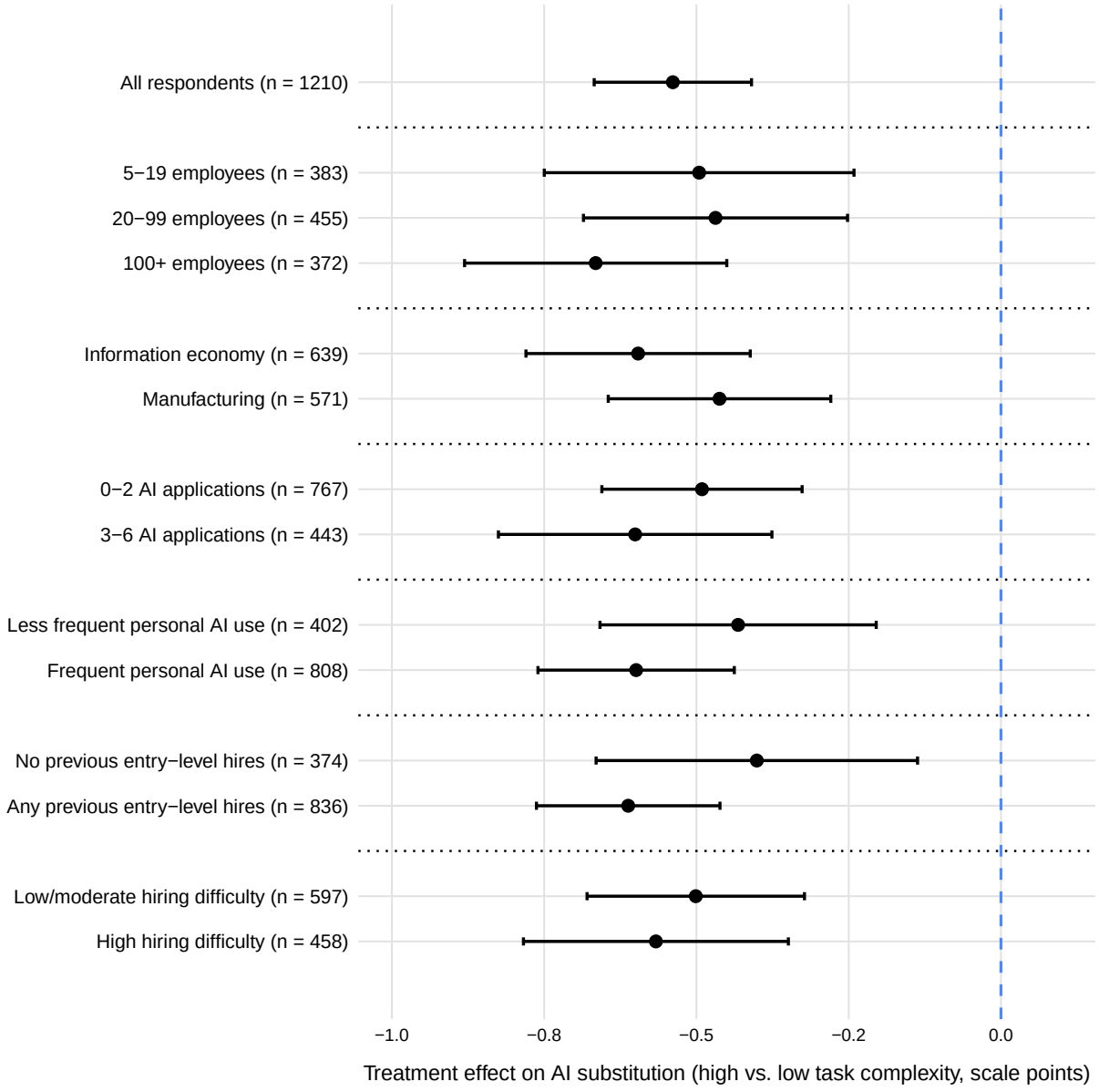


Figure B.2: Treatment-effect heterogeneity in AI substitution

Notes: Points report subgroup-specific OLS estimates of the effect of high rather than low task complexity on willingness to use AI instead of hiring an entry-level employee. Horizontal lines indicate 95% confidence intervals based on HC3 heteroskedasticity-robust standard errors. The analysis uses the complete-case sample of the fully adjusted main specification. Sample sizes are reported separately for each subgroup. Formal interaction tests provide no statistically significant evidence of treatment-effect heterogeneity across the displayed moderators.

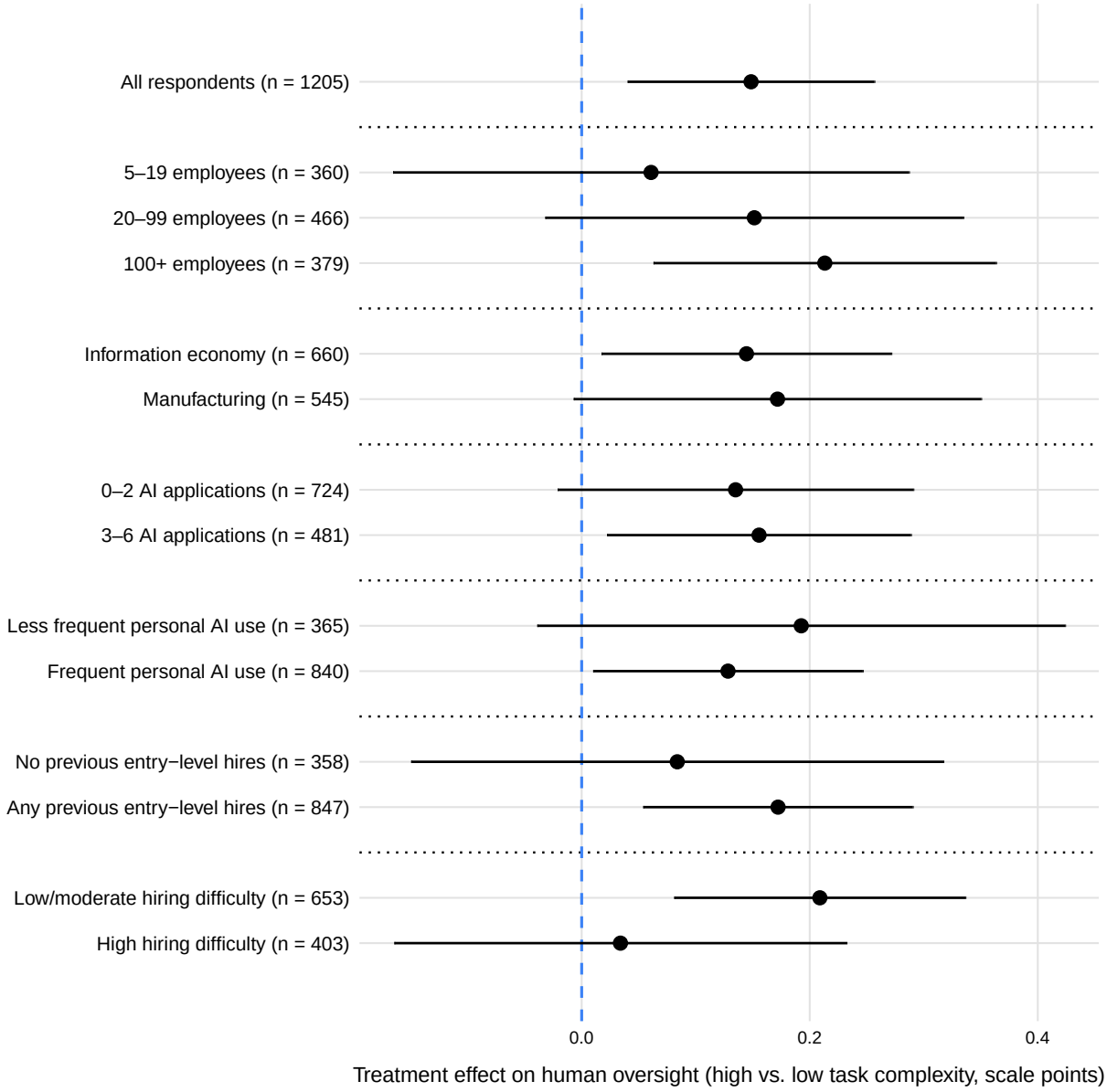


Figure B.3: Treatment-effect heterogeneity in the importance of human oversight

Notes: Points report subgroup-specific OLS estimates of the effect of high rather than low task complexity on the reported importance of human oversight. Horizontal lines indicate 95% confidence intervals based on HC3 heteroskedasticity-robust standard errors. The analysis uses the complete-case sample of the fully adjusted main specification. Sample sizes are reported separately for each subgroup. Formal interaction tests provide no statistically significant evidence of treatment-effect heterogeneity across the displayed moderators.

B.6 Analytical Samples

Table B.6: Analytical samples and treatment allocation

Sample	N	Share of eligible sample (%)	High complexity	Low complexity
Eligible firm sample	1435	100.0	705	730
Valid AI-substitution outcome	1371	95.5	672	699
Full-control sample: AI substitution	1210	84.3	584	626
Valid human-oversight outcome	1359	94.7	673	686
Full-control sample: human oversight	1205	84.0	588	617
Valid values for both primary outcomes	1309	91.2	647	662
Valid entry-level hiring expectation	1247	86.9	618	629
Valid experienced hiring expectation	1291	90.0	645	646
Valid hiring expectations for both groups	1230	85.7	612	618

Notes: The eligible firm sample comprises firms with at least five employees after application of the sample restriction. Shares are calculated relative to the eligible sample of 1,435 firms. Valid outcome samples include respondents providing a value on the respective five-point response scale. Full-control samples contain complete information for all covariates included in the fully adjusted specifications. Hiring-expectation samples exclude firms reporting no general need for the respective employee group. High and low complexity refer to the two randomly assigned treatment conditions.