No Great Equalizer: Experimental Evidence on Productivity Effects of Generative AI Use in the UK Labor Market

Matthias Haslberger*
University of St. Gallen
matthias.haslberger@unisg.ch

Jane Gingrich
University of Oxford
jane.gingrich@spi.ox.ac.uk

Jasmine Bhatia
Birkbeck, University of London
jasmine.bhatia@bbk.ac.uk

March 23, 2025

Abstract

An emerging consensus holds that generative artificial intelligence (AI) equalizes workers' performance within tasks, reducing productivity differences across workers. Existing research has largely studied productivity within single occupational groups and task structures. Whether this equalizing pattern generalizes to the labor market at large remains unclear. Observed performance equalization within groups of workers is compatible with both increasing and decreasing inequality between groups. To distinguish these outcomes, we conducted a large pre-registered online experiment with a sample of the UK working age population which randomly assigned participants to treatments that encouraged or discouraged the use of ChatGPT and then asked them to complete a set of realistic work tasks. We find that ChatGPT use increased productivity in all tasks, with greater benefits observed in more complex and less ambiguous tasks. However, compression effects between tasks were limited. Moreover, ChatGPT use did not affect productivity differentials between gender, age, educational or occupational groups.

Keywords: artificial intelligence, labor market, productivity, inequality, experiment, technological change

*Corresponding author

1 Introduction

An emerging body of work on the labor market implications of generative AI tools such as the ChatGPT family of large language models (LLMs) points to two stylized consequences: they increase productivity in non-routine cognitive tasks, and these gains mostly accrue to those with lower levels of productivity. When scaled up, these arguments suggest both an asymmetrically disruptive effect of generative AI on workers — it will disproportionately impact more educated and professional workers who engage in non-routine cognitive tasks — and an equalizing aggregate effect — it will compress differences among workers. Indeed, these claims have combined into a popular narrative of generative AI threatening white collar jobs. "Decades after automation began taking and transforming manufacturing jobs, artificial intelligence is coming for the higher-ups in the corporate office", writes the Wall Street Journal (Smith, 2024), while the IMF warns about AI "potentially amplifying job losses in cognitive occupations" (Brollo et al., 2024, 1). According to Autor (2024, 2), "AI used well can assist with restoring the middle-skill, middle-class heart of the . . . labor market that has been hollowed out by automation and globalization," suggesting that generative AI could be an equalizing force.

The experimental studies that underpin this narrative show that generative AI both increases and homogenizes individual-level productivity for specific tasks and among specific white-collar occupational groups. For example, college-educated professionals specializing in writing tasks (Noy and Zhang, 2023), management consultants (Dell'Acqua et al., 2023), software developers (Peng et al., 2023), and call center employees (Brynjolfsson, Li and Raymond, 2025) all become more productive when using generative AI, with the least productive workers experiencing the greatest increase in performance. However, we know little about how these findings generalize across different tasks or occupations and a broader range of skills.

We argue that there are two reasons for caution in accepting the emerging narrative of AI as an equalizing force: first, we do not know for which tasks generative AI most enhances productivity and how these tasks map onto occupations; second, given that socio-demographic factors (e.g. age, education, gender) are correlated with occupational selection, studies that examine a single occupational group make it difficult to distinguish effects that are bounded to particular demographic subgroups from those that cross divides. Variation in task characteristics and average AI skills between socio-demographic groups could reduce aggregate compression effects. The early research showing productivity equalization in specific tasks and subsets of workers leaves open critical questions about the aggregate effects of generative AI.

To address these questions, our study uses an experimental design that varies task structures across a broader cross-section of the public. In the first study of this kind, we recruited a sample of 1,041 working-age UK residents with a ChatGPT account and randomly allocated them to a treatment group that is incentivized

to use ChatGPT and a control group. We then asked them to perform three realistic text-based work tasks of increasing complexity and evaluated the quality of their answers and the time they took to answer. This approach allows us to study the effects of AI use on productivity not only within, but also between a) tasks and b) socio-demographic groups. Our findings suggest that the size of productivity increases depends on the complexity and ambiguity of the tasks at hand, limiting compression effects between tasks. Moreover, ChatGPT use did not materially affect performance inequality between socio-demographic groups. Thus, in a pattern reminiscent of Simpson's paradox, generative AI tools at their current level of capability appear to largely replicate existing inequalities between groups, even as they equalize performance within groups. Our findings call into question the emerging narrative of generative AI as an inherently equalizing force and highlight the need for further research into how the uses of this novel technology affect worker productivity.

2 Background

Two dominant approaches exist to analyzing the effects of generative AI on work: a task-focused lens and a skill-focused lens. While these perspectives are related, they highlight different mechanisms linking technology to workers' performance. A task-based lens focuses on the complementarity between generative AI and particular components of work — i.e. the tasks — that AI is most useful for. The skill-based lens shifts our attention to the complementarities between individual skills and AI tools. Those emphasizing tasks point to the utility of generative AI for supporting non-routine cognitive work while those emphasizing skills point to its utility in narrowing performance gaps between those with higher and lower baseline skills. Both perspectives are based on the routine-biased and skill-biased technological change models, respectively (Acemoglu and Autor, 2011).[1] If AI affects the tasks traditionally associated with higher paid white-collar office work and compresses differences among skilled workers, it could have aggregate equalizing effects.

However, we argue that moving from claims about task complementarity and skill compression to these aggregate effects relies on critical assumptions about cross-group differences, which existing studies do not always test. To show this point, we first simulate different skill- and task-based effects and show that compression effects within tasks are compatible with quite distinct aggregate effects, generating some observational uncertainty across approaches. Note, we focus on complementarity without taking into account possible substitution. To what extent generative AI will lead to outright substitution of workers is still hotly debated and beyond the scope of this study (Agrawal, Gans and Goldfarb, 2019; Korinek and Stiglitz, 2019; Gmyrek, Berg and Bescond, 2023; Frey and Osborne, 2023; Green, 2024).

[1]There is a critique that the task-based model is irrelevant to AI since the commercial adoption of AI will require system change rather than changes to job tasks (Bresnahan, 2024; Agrawal, Gans and Goldfarb, 2022). While there is likely truth to this in the long run, for the understanding of current AI capabilities the task-based approach is highly relevant.

In our baseline (pre-AI) model, individual productivity can be expressed as a function of tasks, with average productivity decreasing as tasks become more complex, and as a function of individual skills, with average productivity increasing as workers become more skilled. To model how AI affects aggregate productivity, we first consider how generative AI may change the productivity of individuals for a given task, i.e. varying only across skill level, the approach of most existing work on the productivity effects of generative AI. We then model how observed effects within specific task-skill constellations are compatible with different overall productivity effects.

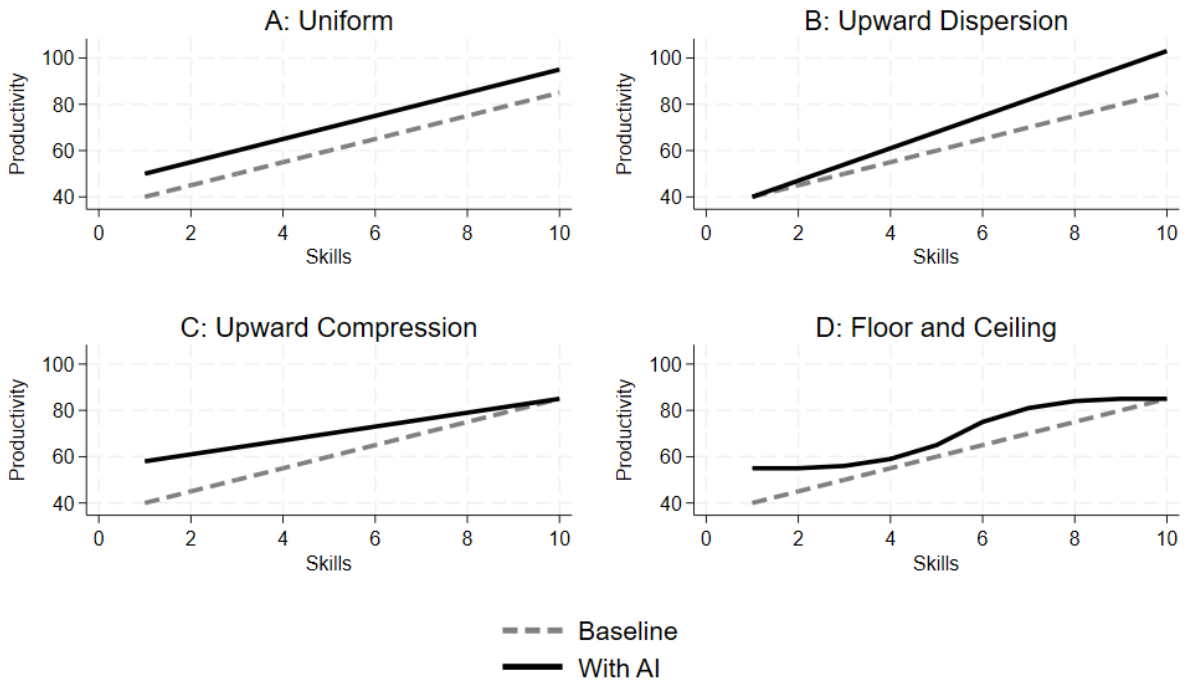
Figure 1 outlines the first step, showing theoretically four different ways how AI may affect productivity across skill groups while holding task structure constant. First, it may be uniformly productivity enhancing, thus maintaining existing skill gradients (A); it may complement existing skills and thus increasing skill gradients (B); it may compress skill differences in productivity reducing variation (C); or it may create floor and ceiling effects that compress skill differences in productivity non-linearly (D, essentially a special case of upward compression).[2] A number of studies of generative AI on productivity look to distinguish these scenarios empirically, with relatively strong emerging evidence that favors scenario C. One study shows that college educated workers doing mid-level professional writing tasks experience an increase in overall productivity and a reduction in inequality between workers when using ChatGPT (Noy and Zhang, 2023). Other studies find a similar compression of productivity due to AI use among highly skilled management consultants (Dell'Acqua et al., 2023), programmers using GitHub Copilot (Peng et al., 2023), and call center workers (Brynjolfsson, Li and Raymond, 2025). Thus, there is strong evidence that those with higher baseline levels of productivity benefit less from AI use, while less productive workers gain more, leading to upward compression within tasks and occupations. However, these studies do not tell us what happens between tasks and occupations, since they draw from a specific point in the task distribution and their participants represent only a limited subset of the overall skill distribution.

To see why, consider Figure 2. It illustrates both within- and between-task effects of AI on the productivity of workers in tasks of different levels of complexity. In the baseline scenario (no AI), average performance decreases as tasks become more complex. The variation in individual performance, which is normally distributed around the mean for each task, is relatively high and increases with complexity. The figure compares the baseline to two possible scenarios that are both compatible with Panel C of Figure 1: both show upward compression, that is, an increase in average performance and reduced variation in performance within tasks, as less productive workers benefit the most.

Where the two alternative scenarios differ is in the stylized effects of AI use on productivity differentials between tasks. Crucially, we see that upward compression within tasks is in principle consistent with both

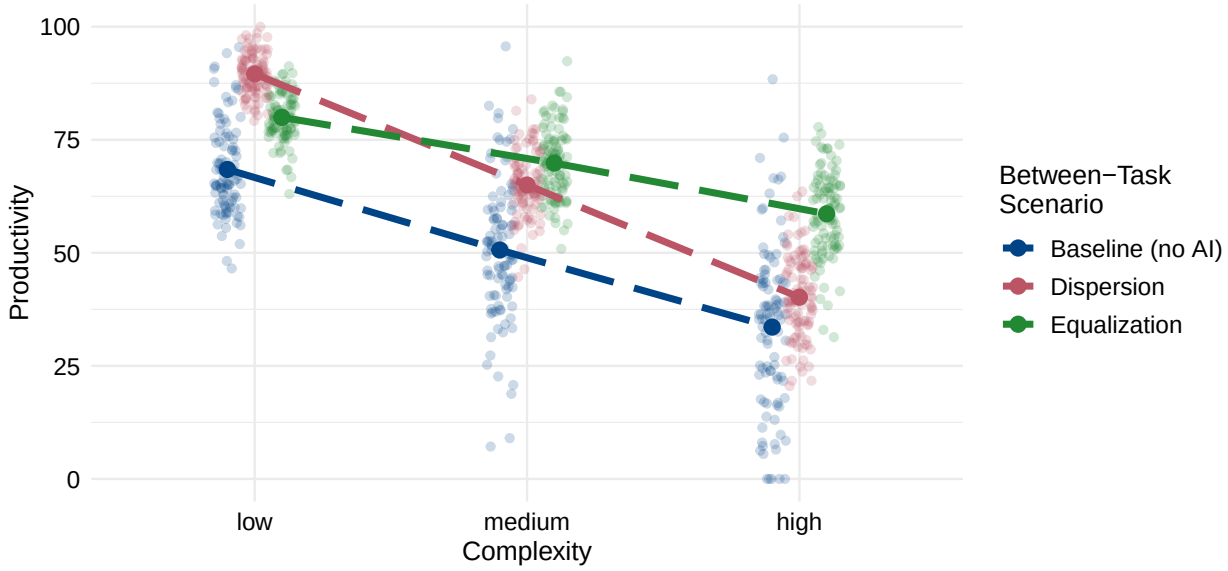
[2]Other non-linear effects, such as winner-take-all effects are also possible.

Figure 1: Stylized within-task skill-based effects



Note: The figure shows four different scenarios linking baseline skills and AI use to performance in a given task.

Figure 2: Simulated within- and between-task dynamics



Note: This figure illustrates how the within-task effects of AI use that have been found in previous studies (upward compression) are compatible with different between-task dynamics (dispersion and equalization).

dispersion and equalization between tasks. The dispersion scenario shows a larger average productivity improvement in the less complex tasks, resulting in a fanning-out of the task-productivity distribution. The equalization scenario shows a larger average productivity increase in more complex tasks, which compresses productivity differentials between tasks compared to the baseline scenario. This pattern resembles Simpson's paradox (Simpson, 1951).

Given that occupations are bundles of tasks (Acemoglu and Autor, 2011) that tend to cluster around certain points of the task distribution (say, e.g., management consultants mostly perform high-complexity tasks, while call center work mostly consists of medium-complexity tasks), these scenarios carry very different implications for productivity differentials between occupations. To understand how AI will affect overall inequalities, it is therefore imperative to consider a wider range of the task and skill distributions and to understand which kinds of tasks AI is most useful for. As Figure 2 highlights, existing research on within-occupation dynamics is consistent with both an equalization and a fanning-out of productivity differences between tasks and occupations. Since workers' earnings depend fundamentally on their productivity (Autor, 2014), this has obvious implications for economic inequalities.

To summarize, existing research leaves important questions regarding the aggregate productivity effects of generative AI unanswered. The narrative of generative AI as inherently equalizing is based on generalizations from occupation-task specific studies, which are compatible with different types of overall outcomes. To establish firmer foundations for empirical expectations regarding aggregate productivity effects, our study looks to both vary the characteristics of the tasks that are exposed to AI and of the individuals who use the technology.

2.1 Task-Based Dynamics: What Is AI Most Useful For?

Work on the effects of ICT and robotics on tasks, traditionally argued that tasks that are hard to specify ex-ante, such as those drawing on tacit knowledge or humans' unique cognitive capacities (e.g. creativity, judgment), are difficult to replace with machines, insulating this work from replacement (Autor, Levy and Murnane, 2003; Autor, 2015). As the technology has evolved, AI can increasingly perform tasks previously restricted to humans, above all non-routine cognitive tasks that require abstract reasoning, prediction, judgment, or creativity (Felten, Raj and Seamans, 2023; Eloundou et al., 2023; Bohren, Hakimov and Lalive, 2024), and even scientific discovery (Lu et al., 2024; Schmidgall et al., 2025; Han, 2024). However, like previous technological advances, generative AI is useful for some tasks but not for others. Early frontier models have struggled with some fairly simple logic and mathematical problems (Berglund et al., 2023; Frieder et al., 2023), and continue to suffer from hallucinations (Gosmar and Dahl, 2025).

Which tasks are most amenable to AI-generated productivity improvements remains unclear. In what Autor (2022, 18) calls 'Polyani's revenge', "computers now know more than they can tell us," at least in some domains.[3] Indeed, one influential study speaks of a "jagged frontier" of AI capabilities (Dell'Acqua et al., 2023), highlighting the unequal pace of progress that turns the capabilities of generative AI into a moving target.

To systematize this dynamic situation, we conceptualize task-based complementary with generative AI based on task complexity, defined as the difficulty for a human of average cognitive abilities to perform the tasks without technological help. This conceptualization can replace the 'non-routine—routine' distinction of the legacy task-based model.[4] Generative AI models have been described as a general-purpose technology which contains "sparks of artificial general intelligence" or "decision machines" (Agrawal, Gans and Goldfarb, 2022, 2019; Eloundou et al., 2023; Bubeck et al., 2023; Goldfarb, Taska and Teodoridis, 2023). Thus, the routineness of a task is less relevant for generative AI, and AI achieves the performance levels of a competent — though not exceptional — human in an increasing number of domains. The potential of generative AI to improve task productivity is fundamentally determined by the gap between average and competent human performance — which is largely a function of how complex the task is for the average person without the help of technology.[5]

Viewed from this perspective, generative AI should reduce variation between tasks. Most people can already perform simple tasks at a competent level and in a reasonable amount of time, leaving limited space for AI to improve productivity. By contrast, more complex tasks, where a larger proportion of individuals struggle, provide ample room for AI to improve performance, both in terms of output quality and time needed to produce the output. By increasing floor-level performance in such tasks, AI can have a large positive effect on mean performance in complex tasks. Collectively, these dynamics should reduce the performance differential between more and less complex tasks. Put differently, in a world with partial AI adoption, the gap between AI users and non-users should increase as task complexity rises. This reasoning, extrapolating from within- to between-task dynamics, underpins the predictions that generative AI will be an equalizing force.

To sum up, one of the key qualities of generative AI is its versatility: many tasks that were previously insulated from technological change are now within its reach; exposure is moving up the skill ladder (Gallego and Kurer, 2022; Knotz, 2025). There are good reasons to believe that the complexity of a task is crucial

[3] 'Polyani's paradox' was that "we know more than we can tell," making it impossible to automate seemingly simple activities that rely on tacit knowledge (Autor, 2015, 11).

[4] The task-based model also distinguishes between manual and cognitive tasks (Autor, Levy and Murnane, 2003), but at least until now generative AI applications focus firmly on the cognitive domain.

[5] With future progress in AI capabilities, such as the recent emergence of reasoning models like ChatGPT-o1, this may of course change. Thus, we do not claim that complexity is the only task characteristic that matters, but we argue that it is a useful starting point for a systematic attempt to understand the productivity effects of AI.

to how much AI improves performance. Previous research focusing on within-occupation dynamics suggests that less able workers experience the greatest benefits. Relying on this insight, we argue that more complex tasks, where a larger proportion of individuals struggle on their own, provide more room for AI to improve performance. While we expect AI to improve performance in all tasks, this suggests that the performance gap between AI users and non-users should grow with task complexity. In this case, generative AI could reduce productivity differentials not just within, but also between tasks.

2.2 Skill-Based Dynamics: Who Is AI Most Useful For?

Another key question to which existing research so far provides an incomplete answer is, who is AI most useful for? Longstanding work suggests that novel technology largely complements high skills (Goldin and Katz, 2008). However, the relationship between AI technology and skills may be very different. The emerging consensus that less able workers within a group benefit the most indicates a potential departure from the pattern observed in digitalization processes so far. Despite much commentary, potential group differences in benefits from generative AI have so far received little attention in empirical research. Yet, there are good reasons to expect AI to exacerbate inequalities along socio-demographic lines. A useful angle to approach this issue is to ask, who develops the necessary expertise to effectively use AI and are there differences between socio-demographic groups that reinforce or alleviate existing labor market inequalities? We argue that AI skills are developed through exposure, as interacting with the technology helps workers build skills. It follows from this that more experienced users should be more proficient and benefit more from using AI. Group differences in AI use are therefore key to understanding differences in benefits from AI. As existing research on 'digital divides' shows, younger, male, more educated, and white-collar individuals are more likely to use digital technologies (OECD, 2018).

Early evidence suggests that this pattern also applies to generative AI. Representative studies from the US and Denmark show that younger, more educated individuals and those working in white-collar occupations are more likely to use generative AI (Humlum and Vestergaard, 2024; Bick, Blandin and Deming, 2024). Most remarkable, however, is the gender gap in ChatGPT use, which amounts to 9 percentage points in the US study and 20 percentage points in the Danish study (Humlum and Vestergaard, 2024; Bick, Blandin and Deming, 2024). This reflects research which shows that in general, women tend to be more skeptical of technology use at work (Gingrich and Kuo, 2022; Borwein et al., 2024), which could prevent them from taking full advantage of the opportunities offered by AI. Indeed, a study with Norwegian economics and business students finds male participants to be more likely to use ChatGPT and consequently more skilled at writing successful prompts (Carvajal, Franco and Isaksson, 2024). Existing research is therefore clear

about the descriptive patterns, and at least one study also substantiates the hypothesized link between AI use and skill development (Carvajal, Franco and Isaksson, 2024). Therefore, differences in AI adoption could result in younger, male, more educated, and white-collar workers obtaining a head start and being able to use AI more effectively than other groups. As in many cases these groups are already advantaged in the labor market, this highlights a potential disequalizing effect of generative AI.[6]

2.3 Hypotheses

To summarize, our theoretical discussion highlights important gaps in the emerging evidence on the labor market effects of generative AI. While a number of papers have convincingly argued that generative AI reduces performance differentials within occupations and tasks, we lack quantitative evidence on the effect of generative AI across a wider range of the task and skill distribution. We posit that AI should increase productivity more strongly in more complex tasks, where many workers otherwise struggle. Second, we follow the existing literature and expect that generative AI has an equalizing effect within tasks. Finally, despite large inequalities in the labor market, there is little evidence on how the effects of AI use on performance differ between socio-demographic groups. Reflecting differences in technology adoption, we expect that older, female, less educated, and blue-collar respondents lag behind in developing AI expertise, and so on average benefit less from using it. These are likely to lead to ongoing *overall differences* in performance, alongside *within-group compression*.

To formally test these claims, we designed an experimental study that was pre-registered on OSF. Our concrete expectations are outlined below.[7]. Hypotheses 1c and 3 are the main novel contributions of our study, while H1a, H1b, and H2 aim to replicate and generalize existing findings.

H1a: Respondents who use ChatGPT will outperform respondents who do not use ChatGPT.

H1b: Respondents who use ChatGPT will take less time to complete the tasks than respondents who do not use ChatGPT.

H1c: The productivity differential between ChatGPT users and non-users will increase in accordance with the complexity of the tasks.

[6]We focus on complementarity, where workers becoming more productive is good for them. As we discuss in more detail below, in practice, substitution may lead to different outcomes, as some sectors may experience employment losses if demand growth does not keep up with productivity growth. Some evidence suggests that this has already happened to freelance knowledge workers (Hui, Reshef and Zhou, 2024; Qiao, Rui and Xiong, 2025).

[7]The pre-analysis plan is available under https://osf.io/dgh2u/?view_only=f30331b363c64ab883d9f76ae0565c68. Deviations from the pre-analysis plan are detailed in Appendix B. Most importantly, we estimate LATE instead of ITT effects in the main text (ITT estimates are in the Supplementary Materials). To reflect the focus on AI use rather than exposure, we slightly reformulated the hypotheses.

H2: Exposure to ChatGPT reduces within-task variation in performance compared to the control group.

H3: a) Female/ b) Older/ c) Less educated/ d) Non-professional respondents are less likely to use AI, and benefit less from using it.

3 Methods

3.1 Experimental Design

Prior to data collection, our study received ethics clearance from the University of Oxford Department of Social Policy and Intervention's Ethics Committee. For our pre-registered online experiment, we recruited 1,041 respondents who had ChatGPT accounts through the survey company YouGov.[8] In contrast to other studies of the effects of ChatGPT (Noy and Zhang, 2023; Dell'Acqua et al., 2023), our sample is larger and approximates the demographic composition of the UK working-age population. The sample was split into a treatment group (N = 504) which was encouraged to use ChatGPT and a control group (N = 537) which was instructed to complete the survey without using ChatGPT.[9] Fieldwork for the survey took place from 19th – 28th July 2023, following a pre-test in early July 2023.[10] The median time to complete the survey was approximately 28 minutes. To increase effort, participants were informed before undertaking the survey that they would be compensated at twice the normal YouGov rate for a survey of this length. In addition, we offered a performance-based reward equal to 50% of the participation reward to the top 10% of respondents in each group. After some self-assessment questions, participants were asked to complete three short text-based tasks of increasing complexity, which form the core of the study. YouGov provided demographic information, including age, sex, education, occupation, income, and other variables.

3.2 Tasks

Study participants performed three tasks of varying complexity (see Table 1) while being instructed to use ChatGPT (treatment group) or not use ChatGPT (control group). We evaluated productivity based on two measures: time to complete the task and answer quality according to pre-specified grading criteria.

[8]It was necessary to sample from ChatGPT users since at the time it was not possible to embed ChatGPT directly in the survey for technical and privacy reasons.

[9]This means that we can directly observe only treatment assignment, not compliance. We describe below how we address this in our estimations.

[10]Extensive documentation of the survey, including sample characteristics (Table S7) balance tests (Table S8), and further descriptives, as well as the tasks and grading scheme (Appendix E), is provided in the supplementary materials. Despite their best efforts, YouGov could not at the time provide a fully representative sample of ChatGPT users. For example, men are overrepresented in the sample (see Table S7). However, we find the experimental results to be unaffected by the inclusion of sample weights or demographic controls (see Tables S23 and S24).

Table 1: Overview of the tasks

Task	Description	Skills tested	Complexity
Email	Improve email addressing a hypothetical workplace dispute	Professional communication, diligence, knowledge of spelling and grammar	Low
Assessment	Evaluate persuasiveness of two texts presenting opposing views	Appraisal of conflicting information, persuasive reasoning	Medium
Comprehension	Answer questions about a complex text	Dealing with complex information, precision	High

Note: Please refer to Appendix E for more details.

In the first task, participants were presented with an email addressing a hypothetical workplace dispute and were asked to make any improvements they deemed appropriate. The email included deliberate grammatical and spelling errors and was written in a harsh, unprofessional tone. The evaluation criteria consider whether the errors have been corrected and whether the tone of the email has been made more constructive. In the second task, respondents evaluated the persuasiveness of two texts of about 500 words, each presenting opposing views on the merits of a universal basic income. The answers were graded based on whether they provided a well-reasoned and comprehensive assessment. It was immaterial which text was chosen as the more persuasive one; both texts could reasonably be considered more persuasive. The quality of the argumentation was decisive. In the final task we asked participants to answer three short questions about a complex text on the economic history of Ireland.[11] The grading criteria evaluate whether respondents correctly distilled the relevant information from the text. Throughout the paper, we refer to the three tasks as the email task, assessment task, and comprehension task, respectively. The three tasks are designed to increase in complexity for a human without technological assistance,[12] and are of a sufficiently general nature so that not only knowledge workers are likely to be familiar with the kind of task, even if not the substance.[13] This design allows us to investigate the impact of AI use for tasks of varying complexity in a sample of the UK working-age population. Compared to earlier studies, our tasks — like our sample — are taken from a wider distribution, allowing for a greater degree of generalizability. However, the focus on text-based tasks entails that we ignore other use cases where generative AI has proved useful, such as coding tasks or creative work. This should be kept in mind when interpreting our results.

[11] The text and questions are adapted from a publicly available LSAT practice test (Law School Admission Council, 2007).

[12] We define complexity as the difficulty for a human of average cognitive abilities to perform the tasks without technological help.

[13] Nevertheless, with increasing complexity, an increasing share of respondents report that they rarely or never perform similar tasks, as we show in Figure S3. However, there are no systematic differences between the treatment and control group.

3.3 Statistical Analysis

We obtained two measures of respondents' performance on the tasks. First, we evaluated their answers on a five-point scale using GPT-4 and a detailed grading scheme (a detailed documentation of the tasks and grading scheme is provided in Appendix E). This led to substantial time savings compared to manual coding, illustrating the potential of LLMs to drastically increase the efficiency of coding text data for quantitative analysis (Gilardi, Alizadeh and Kubli, 2023; Calderon, Reichart and Dror, 2025). To assess the reliability of this approach, two of the authors coded a random subset of 100 answers based on the same grading scheme. Average inter-coder agreement between ChatGPT and human coders, as measured by intra-class correlations, was similar to human- human inter-coder agreement (see Table S9), and while human coders tended to give slightly higher scores overall, there was no systematic ChatGPT-human difference in the score difference between the treatment and control group (see Table S10). This alleviates concerns that AI scores might be endogenous, with ChatGPT recognizing AI-generated answers and evaluating them more positively. As a second measure of productivity, we recorded the time (in seconds) respondents spent on each of the tasks. Since some respondents spent a large amount of time on a single task, we top-coded the time measure for each task at 900 seconds to avoid obtaining results that are driven by a small number of respondents who likely became distracted. Our results are robust to choosing a less stringent cut-off and to excluding those respondents.

We estimated the local average treatment effect (LATE) of using generative AI with a two-stage least squares (2SLS) procedure. Thus, we regressed productivity P_i on self-reported compliance with the treatment instructions C_i, instrumented for with treatment assignment T_i. This yields the following set of equations:

$$C_i = \gamma_0 + \gamma_1 T_i + \upsilon_i \tag{1}$$

$$P_i = \alpha + \beta_1 \hat{C}_i + \epsilon_i \tag{2}$$

Where (1) is the first stage and (2) the second stage. In the analyses of group differences, we added a moderator M_i to the equations. The 2SLS approach allows us to address concerns that due to our experimental design we were unable to directly monitor compliance with our treatment instructions (Allcott et al., 2020).[14] Since treatment allocation is random and demographic attributes are balanced between the two groups (see Table S8), the inclusion of control variables was not necessary. Nevertheless, supplementary

[14]Here we deviate from the pre-analysis plan, in which we stated that we would use linear models and estimate the intention-to-treat (ITT) effect. However, the LATE represents a more appropriate quantity of interest given the experimental setup (Allcott et al., 2020). We report ITT estimates in the supplementary materials (Tables S19 – S22). Appendix D further details the various methods we used to encourage and monitor compliance.

analyses confirm that our results are robust to including demographic controls and weights (see Tables S23 and S24).

4 Results

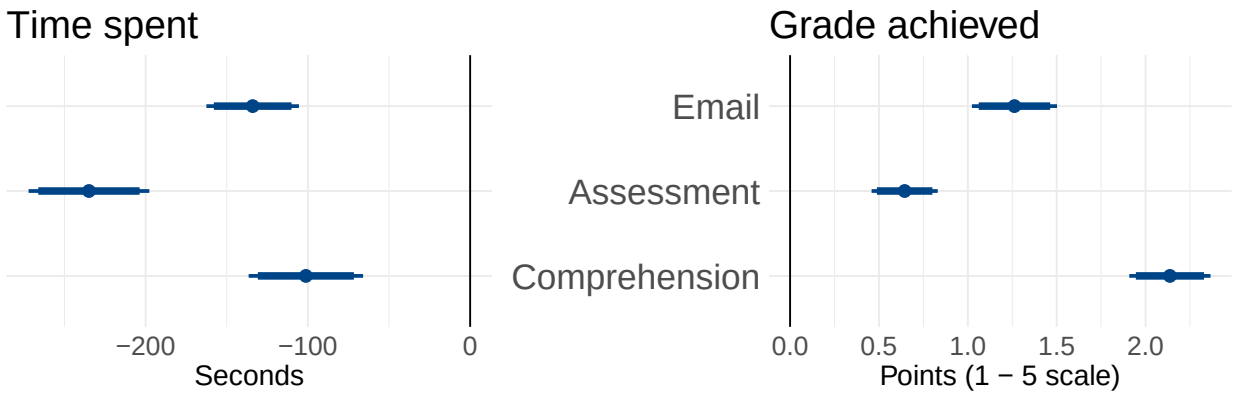
4.1 H1a, H1b, H1c: Overall, Between-, and Within-Task Differences

We find clear evidence that ChatGPT helps respondents give better answers to the tasks in a shorter amount of time (H1a and H1b). Figure 3 shows the LATE of using AI.[15] The difference in time spent and grade achieved between treatment and control group is highly statistically significant ($p < 0.001$) and sizeable for all three tasks. Predicted time savings range between 101 and 235 seconds (0.54 to 1.04 standard deviations (SD)), while performance increases amount to between 0.65 to 2.14 points (0.61 to 1.45 SD). This provides clear evidence for hypotheses 1a and 1b: respondents who used AI tended to perform text-based tasks of varying complexity significantly faster and better. This is in line with evidence from existing studies, which find effects of comparable magnitude (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Brynjolfsson, Li and Raymond, 2025).

One of our contributions to the literature is that we explicitly theorized the relationship between task complexity and task complementarity with AI. However, we find no support for the hypothesis that the return to ChatGPT use increases monotonically with task complexity (H1c). This expectation was based on the reasoning that if less able workers stand to benefit most from using generative AI (Noy and Zhang, 2023; Dell'Acqua et al., 2023), we should see the largest effects in tasks where workers struggle the most on their own. This expectation, however, was not confirmed. As Table S11 shows, respondents in the control group achieved the highest average scores in the email task (3.1), with successively lower grades in the assessment task (2.4) and the comprehension task (2.0), validating our assessment of task complexity. However, AI use increased the average score in the — arguably least complicated — email task by 1.26 points (0.92 SD), but boosted scores in the assessment task by only 0.65 points (0.61 SD). As expected, the return to AI use was greatest in the comprehension task with 2.14 points (1.45 SD). Thus, while AI use led to a substantial and significant increase in answer quality in all tasks, the magnitude of the increase is not simply a function of task complexity.

[15] We present intention-to-treat (ITT) estimates (effect of treatment assignment on outcomes) in the supplementary materials (Tables S19 – S22).

Figure 3: LATE of AI use in different tasks



Note: The figure shows the local average treatment effects estimated using a 2SLS approach, with 90% and 95% confidence intervals. N = 1,041. Full model output in Table S1.

4.2 H2: No Consistent Effect on Within-Task Variation

We now turn to our second question: did AI reduce overall variation? As outlined above, existing research also suggests an effect of AI use on the variation in performance within tasks. If less able workers benefit the most from AI and if we assume that the relative ranking of individuals is maintained, it follows that the variation in performance should be lower in the treatment group (H2). To test this hypothesis, in Table 2 we calculate the coefficient of variation (CV) as a measure of the relative variability of the grades in each task and group.[16] Just as between tasks, we see no consistent pattern. Compared to the control group, within-task variation in grades in the treatment group is lower but only marginally statistically significantly different for the email task ($p < 0.1$), significantly higher for the assessment task ($p < 0.001$), and significantly lower for the comprehension task ($p < 0.05$). Thus, variation in answer quality relative to the control group exhibits no clear tendency in our set of three increasingly complex tasks. Even within broad occupational categories, we see no unambiguous pattern (see Table S18). This qualifies the findings in other recent studies which document upwards compression of within-task performance in occupation-specific tasks and samples. Across a wider range of tasks and with a sample drawn from the entire UK working-age population, generative AI use does not universally compress within-task variation even as it improves average performance. The lack of a consistent pattern suggests that rather than on the mere fact that AI was used, whether AI use compresses or disperses the distribution of answer quality depends on the nature of the task.

[16]The coefficient of variation is defined as the ratio of the standard deviation to the mean. Figure S4 additionally shows density plots.

Table 2: Within-task variation in grades, by group

Task	CV Treated	CV Control	F-statistic	p-value
Email	36.227	38.344	1.120	0.098
Assessment	41.525	35.552	1.364	0.000
Comprehension	46.412	50.028	1.162	0.044

Note: CV: coefficient of variation.

4.3 Explaining Our Findings: The Roles of Task Ambiguity and Time-Performance Trade-Offs

The finding that the benefits of AI use do not monotonically increase with task complexity — contra H1c — and that AI use has no straightforward relationship with within-task variation in performance — contra H2 — requires an explanation. We present evidence for two potential mechanisms, with different implications for further research and organizational practice.

The smaller improvement and the greater variation in answer quality in the treatment group in the assessment task point to a difference in the nature of this task that is independent of its overall complexity. The assessment task required respondents to make a reasoned judgment; there was no right or wrong answer (as we stated explicitly in the instructions). This is a task of medium complexity for humans, as evidenced by the average answer quality in the control group. However, it is characterized by a high degree of ambiguity; conveying helpful instructions to ChatGPT was not straightforward. Hence, while the assessment task is easier than the comprehension task without ChatGPT, it may be easier to prompt ChatGPT to assist with the comprehension task than with the assessment task. Thus, the potential of ChatGPT to improve performance is not just a function of complexity, but also of whether the task is ambiguous or explicit, with straightforward instructions and a 'correct' answer. Therefore, for tasks inside the "jagged frontier" of AI capabilities, generative AI is least useful in tasks that are both simple, meaning most people perform well without help, and ambiguous, meaning that using AI would not be straightforward. By contrast, it offers potential for substantial improvement in tasks that are complex, yet explicit, such as our comprehension task. In simple and explicit (email task), or complex but ambiguous tasks (assessment task), we expect AI to be moderately useful — in the former because room for improvement is limited, in the latter because many people cannot use AI to its full potential. We schematize this relationship in Figure S8, including approximate positions of our three tasks.

If our reasoning is correct, this also implies an interaction between task characteristics and worker skills. Where it is necessary to navigate ambiguity, obtaining useful output from the AI may require AI-specific skills. This indicates a certain degree of skill complementarity in using AI for tasks which are characterized by high ambiguity, which may counterbalance the compressing effect of AI in other tasks. The significantly

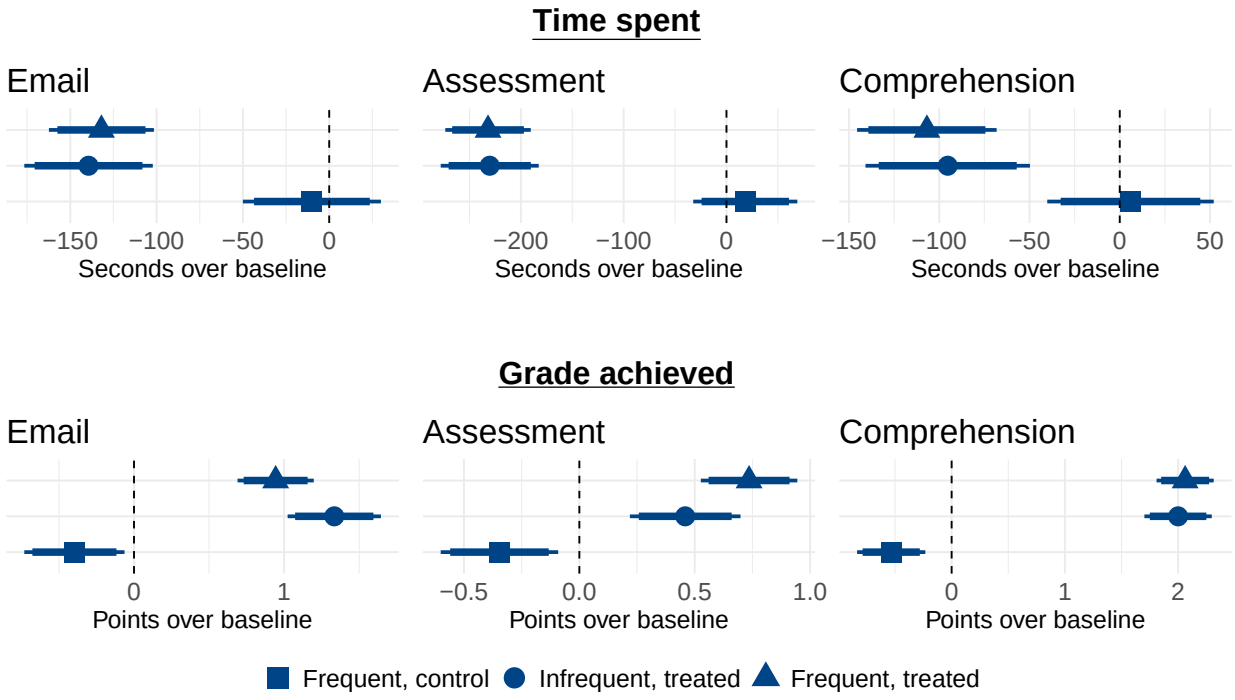
higher coefficient of variation of the treated in the assessment task supports this interpretation. Indeed, there are already real-world examples of such a dynamic in action. For instance, one field experiment with Kenyan entrepreneurs found that, compared to their pre-treatment performance, access to an AI-powered business mentor improved the revenue and profits of high-performing businesses by about 15%, while reducing those of low-performing businesses by about 8% (Otis et al., 2024).

To test this argument explicitly, we move outside of our preregistered hypotheses and look at the above logic. We add an interaction term which indicates whether respondents use AI tools such as ChatGPT frequently (at least once per week) to our 2SLS model. This measure is of course an imperfect proxy for AI-specific skills, but it can provide indicative evidence. Figure 4 shows that there are no significant differences between frequent and infrequent users in either treatment or control group when it comes to time spent on the tasks. In the simple email task, frequent users moreover experienced no greater performance boost than infrequent users. However, in the more ambiguous assessment task and the more complex comprehension task, frequent users saw a much greater improvement with AI, albeit from a significantly lower baseline. In the assessment task, the LATE for frequent users is more than double that for infrequent users, reversing the difference observed in the control group. In the comprehension task, the treatment effect is 30 percent larger for frequent users, erasing the performance differential observed in the control group.[17] This lends empirical support to our argument that AI skills are particularly relevant in more ambiguous tasks. We did not anticipate this qualitative difference between the tasks when pre-registering our hypotheses, which underlines the importance of not only delineating the "jagged frontier" of AI capabilities (Dell'Acqua et al., 2023), but also mapping the territory inside it. Systematizing the determinants of complementarity between tasks and AI will greatly facilitate further research into the consequences of AI. For now, we observe that the hierarchy of task difficulty may differ for humans and AI tools, and posit that the usefulness of generative AI at present depends jointly on the complexity and ambiguity of the task and the skills of the user. The insight that task characteristics influence whether AI-specific skills matter adds further nuance to the empirical finding that "the effects of generative AI are heterogeneous by worker skill type" (Kreitmeir and Raschky, 2024, 3).

Additionally, our findings point to a possible trade-off between performance improvements and time savings. Figure 3 showed that the greater the average improvement in the grades achieved, the smaller the average reduction in time spent. In the assessment task, where AI use improved performance by only 0.61 SD, the time saved amounts to a full 1.04 SD. Conversely, in the comprehension task, where answer quality improved by 1.45 SD, completion time was only 0.54 SD shorter. However, it is impossible to draw

[17] 0.459 points compared to 1.082 points in the assessment task and 2.000 points compared to 2.596 points in the comprehension task, see Table S2.

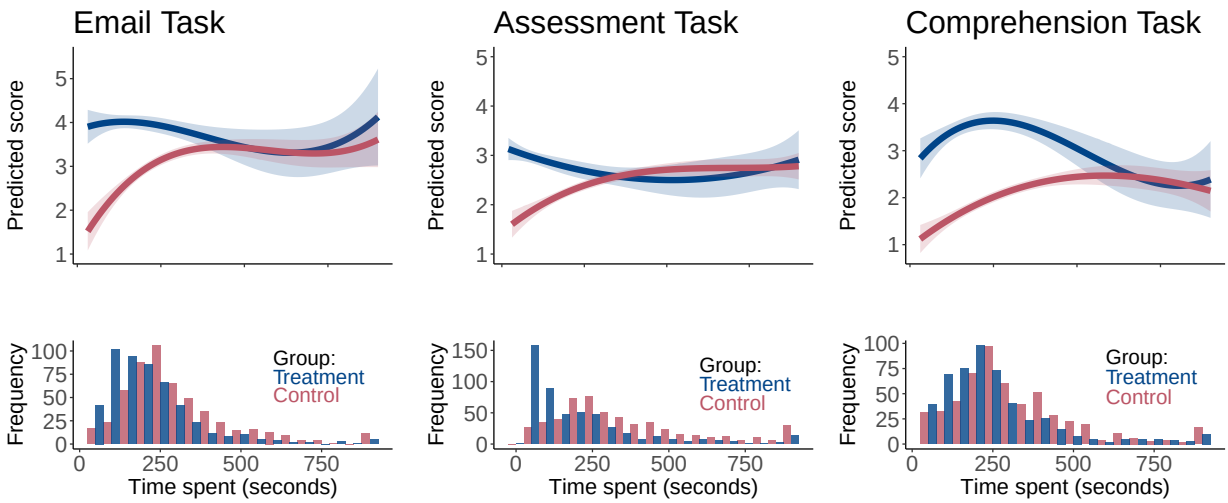
Figure 4: Frequent users benefit more in ambiguous tasks



Note: The figure shows the LATE of using generative AI by frequency of generative AI use, estimated using a 2SLS approach, with 95% confidence intervals. N = 1,041. Full model output in Table S2.

firm conclusions based on this suggestive pattern due to the differences between the tasks: the assessment task involved by far the most reading and took respondents in the control group the longest. Indeed, the individual-level relationship between time spent and performance (see Figure 5) shows a rather different pattern. Respondents in the control group generally see positive, albeit decreasing, returns to time spent, whereas in the treatment group, spending more than approximately 5 minutes on a task is, if anything, associated with a lower score. This is probably partly due to individuals who are less familiar with the technology taking more time to answer the questions, but it also suggests a limited potential for humans to improve over the performance of ChatGPT in our tasks. While we find no evidence for a time-performance trade-off at the individual level, this highlights the general insight that productivity gains from AI allow for adjustments on several margins. Depending on preferences or strategic orientation, individuals and organizations may funnel productivity gains into higher output, higher quality, lower labor input (hours), or any combination thereof.

Figure 5: The time-performance trade-off



Note: This figure shows the relationship between time spent and performance on a task. The predicted values are obtained by fitting a third degree polynomial of time spent on performance.

4.4 H3: Group Differences in AI Use and Productivity

We now investigate group differences in AI use and productivity. Past waves of technological change have often had profound economic and political consequences (Autor, 2015; Mokyr, Vickers and Ziebarth, 2015). Most recently, a large literature shows convincingly that automation and globalization led to concentrated employment losses, exemplified by the decline of factory jobs for middle-skilled white men, which were in

turn followed by political backlash and polarization (Autor, Dorn and Hanson, 2013; Autor et al., 2020). It is therefore imperative to anticipate potential cleavages that may appear in the wake of AI adoption. The previous section indicated that familiarity with generative AI enables workers to use the technology more effectively in more ambiguous tasks. If some demographic groups face barriers to developing AI skills — prompting and evaluating LLM output, and so forth — they might not be able to use the technology effectively. As a consequence, they might struggle to adapt to the attendant changes in work organization and face disproportionate labor market risks. Hence we ask, who uses AI and are there group differences in how effectively they do so?

4.4.1 The "Usual Suspects" Are More Likely to Use AI

We hypothesized that male, younger, and more educated respondents, as well as managers and professionals, would be more likely to use AI, because these groups are generally perceived to be more exposed to cutting-edge technologies and therefore more tech-savvy (OECD, 2018; Humlum and Vestergaard, 2024; Bick, Blandin and Deming, 2024). Bearing in mind that our sample includes only respondents who stated that they have a ChatGPT account, we still find the expected differences in self-reported frequency of use in Table 3. Men, respondents under 40, degree holders, and managers and professionals are between 4 and 9 percentage points more likely to report using ChatGPT at least once per week. We therefore find evidence that patterns of use correspond to prevalent stereotypes regarding digital technologies. Yet, it must be noted that even in our positively selected sample, frequent users are a minority, and user growth has levelled off after the initial surge.[18] Accounting furthermore for purely "recreational" use of AI technologies, it is clear that workplace adoption of generative AI is still in an early stage and that we are in what Agrawal, Gans and Goldfarb (2022, xiii) call "The Between Times — *after* witnessing the power of this technology and *before* its widespread adoption".

Table 3: Frequent users of ChatGPT by demographic characteristics

	ChatGPT Use (%)	
Group	Infrequent	Frequent
Female	76	24
Male	67	33
Under 40	67	33
40 +	74	26
University degree	69	31
Non-university degree	73	27
Professional/managerial occupation	67	33
Other occupation	76	24

Note: Respondents are classified as frequent users if they report using ChatGPT at least once per week and as infrequent users otherwise.

[18] We accessed use statistics on 9 April 2024 on this website.

4.4.2 Who Benefits Most from Using AI?

One crucial question for businesses and policymakers looking into the future, then, is how effectively are people with different levels of affinity towards AI going to use the technology? As Dahlke et al. write, "adoption patterns established in the early stages of technology diffusion can lead to path dependencies" (Dahlke et al., 2024, 1). Thus, if the descriptive patterns of AI use are indicative of AI skills, far from being an equalizer, AI might disproportionately benefit already privileged groups and exacerbate existing labor market inequalities (Lane and Saint-Martin, 2021; Albanesi et al., 2023). To examine H3, which hypothesizes heterogeneous effects by demographic groups, we include interactions of AI use (instrumented by treatment assignment) with sex, age, education, or occupation in the second stage of our estimations. This approach addresses selection into compliance based on demographic characteristics by focusing on the exogenous portion of variation in compliance within each group. Note that the interaction coefficients can be interpreted as the difference in differences of the LATE on the respective groups. Differences in the frequency of use notwithstanding, the benefits of AI use appear to be fairly evenly distributed. Figure 6 shows effects on time spent on the tasks and Figure 7 shows effects on grades achieved for each of the four socio-demographic groups.[19]

Despite the greater initial familiarity of men with ChatGPT and the well-established "gender digital divide" in favor of men (OECD, 2018), we find remarkably similar treatment effects, as seen in Panel A of Figures 6 and 7. While men who used AI experienced a greater reduction in time spent in all tasks compared to the control group, the difference in differences is never statistically significant. Women in the control group achieve significantly higher grades in the email task than men, but in the treatment group there is no such difference. Yet, the difference in differences is not statistically significant. Thus, men do not appear to gain a material productivity advantage from their greater initial familiarity with AI technology, in contrast to, for example, a recent study of Norwegian university students that specifically investigated gender differences in prompting skills (Carvajal, Franco and Isaksson, 2024). Should our finding be replicated and persist in future iterations of the technology, generative AI might contribute to closing the gender gap in digital skills (OECD, 2018) and, in time, the gap in attitudes towards using digital technology (Combet, 2024; Cai, Fan and Du, 2017).

Next, we investigate potential heterogeneity by age, distinguishing between respondents under 40 and those aged 40 years and older. Panel B of Figures 6 and 7 shows that both age groups become significantly more productive on all three tasks when using ChatGPT. Contrary to expectation, the size of the productivity improvement does not appear to differ between age groups; neither for time spent nor for grades achieved do

[19] Full regression tables (including first-stage estimates) for all interaction models are presented in the supplementary materials (Tables S3 – S6).

we find significant interaction effects. Importantly, this is not due to educational expansion: the results are unchanged if we control for degree status. Thus, both younger and older workers become more productive with generative AI in equal measure. Even when distinguishing instead between workers under 50 and those aged 50 years and older, there is no evidence that older workers benefit less from using ChatGPT.[20] However, it is important to bear in mind that our sample likely includes comparatively tech-savvy older workers and our results therefore potentially understate the advantage to younger workers in a fully representative sample.

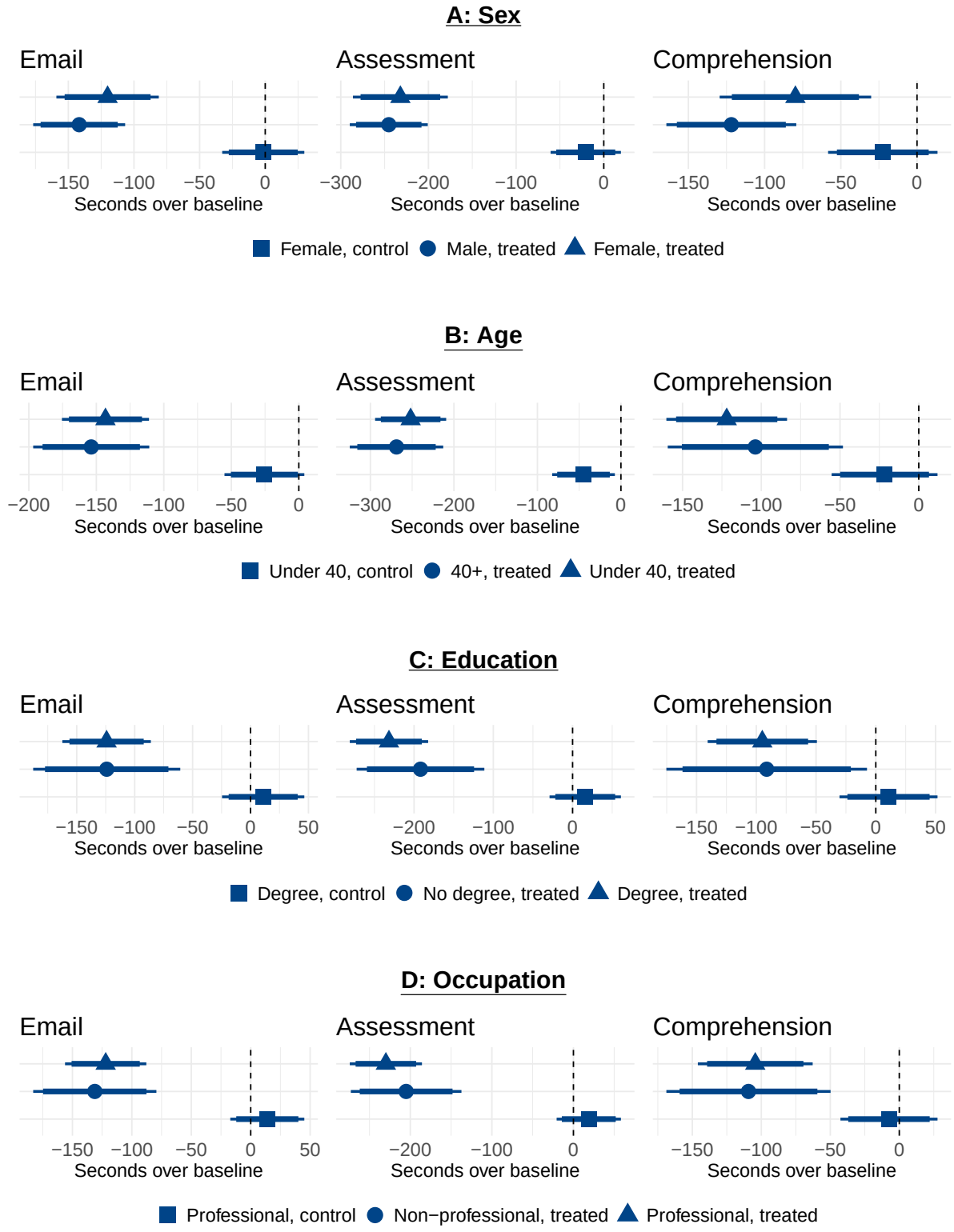
We also find no evidence to suggest that more educated workers or those working in professional or managerial occupations might monopolize the benefits of generative AI. While respondents with a university degree in the control group achieve marginally higher grades on all three tasks, the difference remains short of statistical significance and AI use improves scores and reduces completion times in both education groups by approximately the same amount (see Figures 6 and 7, Panel C). The difference in differences is statistically insignificant in all cases. Similarly, while professionals and managers in the control group achieve marginally higher scores in all three tasks, the treatment only leads to a statistically insignificant reduction in the performance differential (see Figures 6 and 7, Panel D). There are no significant occupational differences in time spent, nor differential treatment effects.

One objection to these — perhaps counterintuitive — null findings might be that we lack the statistical power for interaction analyses. In response, in Appendix C we report extensive power analyses and Bayes factors that validate our conclusions. The power analysis indicates that, at the conventional threshold of 80 percent power and a significance level of $p < 0.05$, we can detect small- to medium-sized interaction effects. By contrast, most of the (insignificant) effects we find are very small, so that even if we were to find them to be significant in a larger sample, they would not be substantively meaningful. Similarly, the Bayes factors show that there is strong evidence for the absence of an interaction between AI use and sex, age, education, and occupation. Thus, the greater initial familiarity with generative AI of male, younger, highly educated, and white-collar respondents does not translate into differential performance improvements that allow these groups to pull away, contradicting our third set of hypotheses. Yet, neither do the findings support a group equalization narrative, as might be extrapolated from existing within-group analyses (Noy and Zhang, 2023; Dell'Acqua et al., 2023). Instead, our results are largely in line with a uniform productivity effect across socio-demographic groups (Panel A of Figure 1).

However, this does not mean that the proliferation of generative AI will have no effect on inequalities in the labor market: there are several other margins where AI can affect group differences. For example, we do find AI to be more useful for some tasks than others, and to the extent that different socio-demographic

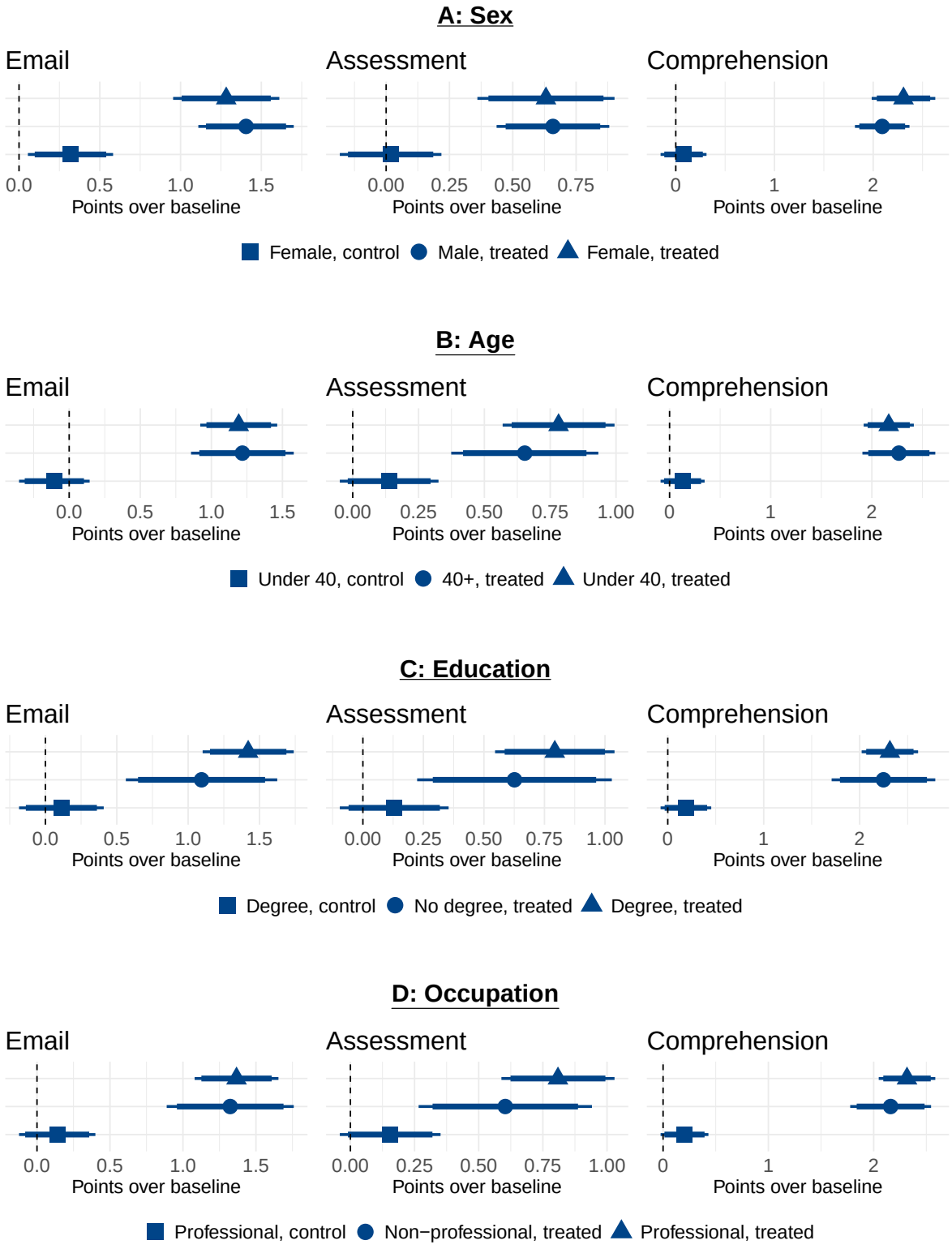
[20]The first-stage regressions in this analysis indicate that workers aged 50 and older in the treatment group are less likely to actually use AI (see Table S25). Significant heterogeneous effects by age in ITT analyses are therefore driven by age differences in compliance with the treatment (see Table S20).

Figure 6: AI use reduces time for completion by similar amounts across socio-demographic groups



Note: The figure shows the LATE of using generative AI on answer quality, by socio-demographic characteristics, with 95% confidence intervals. LATE estimated using a 2SLS approach. Full model output in Tables S3 - S6.

Figure 7: AI use improves answer quality by similar amounts across socio-demographic groups



Note: The figure shows the LATE of using generative AI on answer quality, by socio-demographic characteristics, with 95% confidence intervals. LATE estimated using a 2SLS approach. Full model output in Tables S3 - S6.

groups perform on average different tasks (Grundke, Marcolin and Squicciarini, 2019; Haslberger, 2022; Cortes, Jaimovich and Siu, 2023), AI may narrow or exacerbate inequalities through compositional and demand effects (Bessen, 2019).

Let us take the example of software engineers, a heavily male-dominated and well-paid occupation. Suppose AI tools such as GitHub Copilot lead to a large increase in productivity of individual developers (Peng et al., 2023). Assuming a constant number of software engineers, we would expect to see the unit price of their output decline, demand for, and revenue from, their services increase, and higher salaries to reflect this greater productivity. Due to the composition of the occupational workforce, this would increase the gender wage gap. However, if demand for software engineering work were to fall behind productivity growth, the resulting redundancies would predominantly hit men, reducing the gender employment and wage gaps. Our findings merely show that AI does not confer an inherent productivity advantage on any one socio-demographic group; which of the abovementioned — or other — scenarios prevails is beyond the scope of this study.

5 Discussion

Several studies have found that within occupations, recent iterations of ChatGPT or similar tools equalize worker performance (Noy and Zhang, 2023; Dell'Acqua et al., 2023; Peng et al., 2023; Brynjolfsson, Li and Raymond, 2025). Despite being based on narrow samples, these studies have been used to draw conclusions about the implications of generative AI for the labor market as a whole. This study investigates the emergent narrative that generative AI is likely to be an equalizing force with a survey experiment of a broader sample of the UK working-age population. The participants completed three text-based tasks of varying complexity and ambiguity, with respondents randomized into treatments that encouraged or discouraged the use of ChatGPT. The findings help us better understand what generative AI is most useful for, who benefits most from using it, and how this may affect inequalities at large.

We found that the productivity gains from AI use are substantial both in terms of time savings and quality improvements. However, our expectation that generative AI yields the greatest benefits in more complex tasks was not confirmed. We also found no consistent pattern with regard to within-task variation in performance. This suggests that additional factors besides complexity affect how useful AI is for a task. We propose a taxonomy which places tasks along two axes, ranging from simple to complex and from ambiguous to explicit. AI appears to be most useful in tasks that are complex yet explicit. In simple tasks, the potential for improvement over human performance is limited. In ambiguous tasks, obtaining a meaningful output from the AI is not straightforward, making respondent skill a crucial moderating variable. However,

this perspective requires further empirical investigation while remaining mindful of the rapidly evolving capabilities of generative AI (Mollick, 2024). Yet, mapping and systematizing the "jagged frontier" of AI capabilities is a crucial requirement for studying its impact on businesses and the labor market.

The second part of our analysis dealt with the all-important question who benefits from using AI. Our findings suggest that broad segments of the working-age population are able to use generative AI tools productively and integrate them into their workflows. ChatGPT and similar technologies appear to make workers more productive regardless of their sex, age, education, or occupational background. Again, this may change as future iterations of generative AI continue to evolve.

All in all, our results paint a nuanced picture of the prospective aggregate impacts of generative AI and caution against expectations that AI will be an equalizing force. New inequalities appear most likely to arise from compositional effects or changes in work organization. The centrality of tasks that are complemented by AI and the representation of different demographic groups varies across occupations, and managers have discretion over how AI is incorporated into workflows (Mollick, 2024; Agrawal, Gans and Goldfarb, 2022). In short, AI does not inherently increase or reduce inequalities, and much depends on management and policy choices (Autor, 2024). As the adoption of generative AI in organizations is gathering pace and is expected to have a major impact on labor markets in the years to come (Humlum and Vestergaard, 2024), this insight is both reassuring and a call to action. Skillful management and forward-looking policy can limit the potential negative externalities of AI adoption. As businesses are often slow to adapt, since restructuring around a new technology requires fundamental organizational change (Agrawal, Gans and Goldfarb, 2022; Mollick, 2024; Feigenbaum and Gross, 2024), this opens a window for research to show how this can be achieved.

Before concluding, we hasten to point out that the time and methodological constraints inherent to our online survey design necessitated certain trade-offs. Due to sample size limitations, we could not investigate whether AI training affects people's competence in using AI (Dell'Acqua et al., 2023). Furthermore, due to survey time constraints, we did not collect a baseline measure of task-specific skills which would have allowed for within-individual comparisons (Noy and Zhang, 2023; Dell'Acqua et al., 2023). The same reason also prevented us from including an even wider range of tasks which would have enhanced the generalizability of our findings. Our use of monetary incentives to increase effort may have had a stronger effect on men (Saccardo, Pietrasz and Gneezy, 2018; Samek, 2019). Most importantly, it was not possible at the time to administer the survey to a fully representative sample of the UK working-age population — instead, our sample was limited to individuals in YouGov's panel who already had a ChatGPT account. Our participants are therefore likely to be more open to new technologies than the population at large. Additionally, our study illustrates the effects of generative AI capabilities at a particular moment in time; we acknowledge that future iterations may yield different effects. As such, longitudinal studies which track changes in user

productivity over time would be particularly useful.

To conclude, our study provides causal evidence that generative AI enables workers of all types to be significantly more productive in a range of text-related tasks, with ambiguous effects on between- and within-group inequalities. However, ultimately the impact on organizations and labor markets will depend on the capabilities of future AI tools, as well as the regulatory environment and public attitudes towards AI (Raviv, 2023; Margalit and Raviv, 2023; Schiff, Schiff and Jacobson, 2024; Haslberger, Gingrich and Bhatia, 2024), the reorganization of tasks and workplaces (Autor, 2024; Agrawal, Gans and Goldfarb, 2022; Goos and Savona, 2024), innovations in training practices and skill formation systems (Beane and Anthony, 2024; Beane, 2019; Trajtenberg, 2019), and changes in demand for the products and services made more affordable by AI (Bessen, 2019). Further adding to this "vast uncertainty" (Autor, 2022), our findings challenge the emerging consensus about the equalizing effects of generative AI, and underscore the need for research considering the aggregate effects of AI.

References

Acemoglu, Daron and David Autor. 2011. Skills, tasks and technologies: Implications for employment and earnings. In Handbook of Labor Economics. Vol. 4b Elsevier B.V. pp. 1043–1171. arXiv: 1011.1669v3 ISSN: 15734463.

Agrawal, Ajay, Joshua S. Gans and Avi Goldfarb. 2019. "Exploring the impact of artificial Intelligence: Prediction versus judgment." Information Economics and Policy 47.

Agrawal, Ajay, Joshua S. Gans and Avi Goldfarb. 2022. Power and Prediction: The Disruptive Economics of Artificial Intelligence. Boston, MA: Harvard Business Review Press.

Albanesi, Stefania, Antonio Dias da Silva, Juan F. Jimeno, Ana Lamo and Alena Wabitsch. 2023. "New Technologies and Jobs in Europe." NBER Working Paper Series (31357).

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." American Economic Review 110(3):629–676.

Autor, David. 2014. "Skills, education, and the rise of earnings inequality among the "other 99 percent"." Science 344(6186):843–852.

Autor, David. 2015. "Why Are There Still So Many Jobs? The History and Future of Workplace Automation." Journal of Economic Perspectives 29(3):3–30. ISBN: 9781589017009.

Autor, David. 2022. "The Labor Market Impacts of Technological Change: From Unbridled Enthusiasm to Qualified Optimism to Vast Uncertainty." NBER Working Paper Series (30074).

Autor, David. 2024. "Applying AI to Rebuild Middle Class Jobs." NBER Working Paper Series (32140).

Autor, David, David Dorn and Gordon H. Hanson. 2013. "The China Syndrome: Local Labor Market Effects of Import Competition in the United States." American Economic Review 103(6):2121–2168. arXiv: 1011.1669v3 ISBN: 0002-8282.

Autor, David, David Dorn, Gordon Hanson and Kaveh Majlesi. 2020. "Importing political polarization? The electoral consequences of rising trade exposure." American Economic Review 110(10):3139–3183.

Autor, David, Frank Levy and Richard J. Murnane. 2003. "The Skill Content of Recent Technological Change: An Empirical Exploration." The Quarterly Journal of Economics 118(4):1279–1333.

Beane, Matthew. 2019. "Shadow Learning: Building Robotic Surgical Skill When Approved Means Fail." Administrative Science Quarterly 64(1):87–123.

Beane, Matthew and Callen Anthony. 2024. "Inverted Apprenticeship: How Senior Occupational Members Develop Practical Expertise and Preserve Their Position When New Technologies Arrive." Organization Science 35(2):405–431.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak and Owain Evans. 2023. "The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A".". arXiv:2309.12288 [cs].

Bessen, James. 2019. Artificial Intelligence and Jobs: The Role of Demand. In The Economics of Artificial Intelligence: An Agenda, ed. Ajay Agrawal, Joshua S. Gans and Avi Goldfarb. Chicago: University of Chicago Press pp. 291–307.

Bick, Alexander, Adam Blandin and David Deming. 2024. "The Rapid Adoption of Generative AI." NBER Working Paper Series (32966).

Bohren, Noah, Rustamdjan Hakimov and Rafael Lalive. 2024. "Creative and Strategic Capabilities of Generative AI: Evidence from Large-Scale Experiments.".

Borwein, Sophie, Beatrice Magistro, Peter Loewen, Bart Bonikowski and Blake Lee-Whiting. 2024. "The gender gap in attitudes toward workplace technological change." Socio-Economic Review p. mwae004.

Bresnahan, Timothy. 2024. "What innovation paths for AI to become a GPT?" Journal of Economics & Management Strategy 33(2):305–316.

Brollo, Fernanda, Era Dabla-Norris, Ruud de Mooij, Daniel Garcia-Macia, Tibor Hanappi, Li Liu and Anh D. M. Nguyen. 2024. "Broadening the Gains from Generative AI." IMF Staff Discussion Notes 2024(002):1–43.

Brynjolfsson, Erik, Danielle Li and Lindsey Raymond. 2025. "Generative AI at Work." The Quarterly Journal of Economics p. qjae044.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro and Yi Zhang. 2023. "Sparks of Artificial General Intelligence: Early experiments with GPT-4.". arXiv:2303.12712 [cs].

Cai, Zhihui, Xitao Fan and Jianxia Du. 2017. "Gender and attitudes toward technology use: A meta-analysis." Computers & Education 105:1–13.

Calderon, Nitay, Roi Reichart and Rotem Dror. 2025. "The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs.". arXiv:2501.10970 [cs].

Carvajal, Daniel, Catalina Franco and Siri Isaksson. 2024. "Will Artificial Intelligence Get in the Way of Achieving Gender Equality?" SSRN Electronic Journal .

Combet, Benita. 2024. "Women's aversion to majors that (seemingly) require systemizing skills causes gendered field of study choice." European Sociological Review 40(2):242–257.

Cortes, Guido Matias, Nir Jaimovich and Henry E. Siu. 2023. "The Growing Importance of Social Tasks in High-Paying Occupations: Implications for Sorting." Journal of Human Resources 58(5):1429–1451.

Dahlke, Johannes, Mathias Beck, Jan Kinne, David Lenz, Robert Dehghan, Martin Wörter and Bernd Ebersberger. 2024. "Epidemic effects in the diffusion of emerging digital technologies: evidence from artificial intelligence adoption." Research Policy 53(2):104917.

Dell'Acqua, Fabrizio, Edward McFowland Iii, Ethan Mollick, Hila Lifshitz-Assaf, Katherine C Kellogg, Saran Rajendran, Lisa Krayer, François Candelon and Karim R Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." SSRN Electronic Journal .

Eloundou, Tyna, Sam Manning, Pamela Mishkin and Daniel Rock. 2023. "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models.". arXiv:2303.10130 [cs, econ, q-fin].

Feigenbaum, James and Daniel P. Gross. 2024. "Organizational and Economic Obstacles to Automation: A Cautionary Tale from AT&T in the Twentieth Century." Management Science p. mnsc.2022.01760.

Felten, Edward W., Manav Raj and Robert Seamans. 2023. "Occupational Heterogeneity in Exposure to Generative AI." SSRN Electronic Journal .

Frey, Carl Benedikt and Michael Osborne. 2023. "Generative AI and the Future of Work: A Reappraisal." Brown Journal of World Affairs XXX(1).

Frieder, Simon, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, Alexis Chevalier and Julius Berner. 2023. "Mathematical Capabilities of ChatGPT.". arXiv:2301.13867 [cs].

Gallego, Aina and Thomas Kurer. 2022. "Automation, Digitalization, and Artificial Intelligence in the Workplace: Implications for Political Behavior." Annual Review of Political Science 25(1):463–484.

Gilardi, Fabrizio, Meysam Alizadeh and Maël Kubli. 2023. "ChatGPT outperforms crowd workers for text-annotation tasks." Proceedings of the National Academy of Sciences 120(30):e2305016120.

Gingrich, Jane and Alexander Kuo. 2022. Gender, Technological Risk, and Political Preferences. In Digitalization and the Welfare State, ed. Marius R Busemeyer, Achim Kemmerling, Paul Marx and Kees van Kersbergen. Oxford University Press pp. 157–173.

Gmyrek, Pawel, Janine Berg and David Bescond. 2023. "Generative AI and Jobs: A global analysis of potential effects on job quantity and quality." ILO Working Paper 96.

Goldfarb, Avi, Bledi Taska and Florenta Teodoridis. 2023. "Could machine learning be a general purpose technology? A comparison of emerging technologies using data from online job postings." Research Policy 52(1):104653.

Goldin, Claudia and Lawrence F. Katz. 2008. The Race between Education and Technology. Cambridge, MA: Harvard University Press.

Goos, Maarten and Maria Savona. 2024. "The governance of artificial intelligence: Harnessing opportunities and mitigating challenges." Research Policy 53(3):104928.

Gosmar, Diego and Deborah A. Dahl. 2025. "Hallucination Mitigation using Agentic AI Natural Language-Based Frameworks.". arXiv:2501.13946 [cs].

Green, Andrew. 2024. Artificial intelligence and the changing demand for skills in the labour market. OECD Artificial Intelligence Papers 14. Series: OECD Artificial Intelligence Papers Volume: 14.

Grundke, Robert, Luca Marcolin and Mariagrazia Squicciarini. 2019. Narrowing the gender wage gap in the digital era: the role of skills. In Taking Stock: Data and Evidence on Gender Equality in Digital Access, Skills, and Leadership, ed. Araba Sey and Nancy Hafkin. Macau: United Nations University pp. 282–292.

Han, Sukjin. 2024. "Mining Causality: AI-Assisted Search for Instrumental Variables.". arXiv:2409.14202 [econ].

Haslberger, Matthias. 2022. "Rethinking the measurement of occupational task content." The Economic and Labour Relations Review 33(1):178–199.

Haslberger, Matthias, Jane Gingrich and Jasmine Bhatia. 2024. "Rage Against the Machine? Generative AI Use, Threat Perceptions, and Policy Preferences." SSRN Electronic Journal .

Hoijtink, Herbert, Joris Mulder, Caspar Van Lissa and Xin Gu. 2019. "A tutorial on testing hypotheses using the Bayes factor." Psychological Methods 24(5):539–556.

Hui, Xiang, Oren Reshef and Luofeng Zhou. 2024. "The Short-Term Effects of Generative Artificial Intelligence on Employment: Evidence from an Online Labor Market." Organization Science p. orsc.2023.18441.

Humlum, Anders and Emilie Vestergaard. 2024. "The Adoption of ChatGPT." SSRN Electronic Journal .

Jarosz, Andrew F. and Jennifer Wiley. 2014. "What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors." The Journal of Problem Solving 7(1).

Knotz, Carlo Michael. 2025. "Technological vulnerability among the higher-educated: implications for party preferences." European Political Science Review pp. 1–19.

Korinek, Anton and Joseph E. Stiglitz. 2019. Artificial Intelligence and Its Implications for Income Distribution and Unemployment. In The Economics of Artificial Intelligence: An Agenda, ed. Ajay Agrawal, Joshua S. Gans and Avi Goldfarb. Chicago: University of Chicago Press pp. 349–390.

Kreitmeir, David and Paul A. Raschky. 2024. "The Heterogeneous Productivity Effects of Generative AI.". arXiv:2403.01964 [cs, econ, q-fin].

Lane, Marguerita and Anne Saint-Martin. 2021. "The impact of Artificial Intelligence on the labour market: What do we know so far?" OECD Social, Employment and Migration Working Papers (256).

Law School Admission Council. 2007. "The Official LSAT Preptest.". Form 8LSN75.

Lin, Ruitao and Guosheng Yin. 2015. "Bayes factor and posterior probability: Complementary statistical evidence to p-value." Contemporary Clinical Trials 44:33–35.

Lu, Chris, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune and David Ha. 2024. "The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery.". arXiv:2408.06292 [cs].

Margalit, Yotam and Shir Raviv. 2023. "The Politics of Using AI in Public Policy: Experimental Evidence." SSRN Electronic Journal .

Mokyr, Joel, Chris Vickers and Nicolas L. Ziebarth. 2015. "The History of Technological Anxiety and the Future of Economic Growth: Is This Time Different?" Journal of Economic Perspectives 29(3):31–50.

Mollick, Ethan. 2024. Co-Intelligence: Living and Working with AI. New York: Portfolio/Penguin.

Noy, Shakked and Whitney Zhang. 2023. "Experimental evidence on the productivity effects of generative artificial intelligence." Science (381):187–192.

OECD. 2018. Bridging the Digital Gender Divide - Include, Upskill, Innovate. Paris: OECD.

Otis, Nicholas, Berkeley Haas, Rowan Clarke, Solene Delecourt and David Holtz. 2024. "The Uneven Impact of Generative AI on Entrepreneurial Performance." SSRN Electronic Journal .

Peng, Sida, Eirini Kalliamvakou, Peter Cihon and Mert Demirer. 2023. "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot.". arXiv:2302.06590 [cs].

Qiao, Dandan, Huaxia Rui and Qian Xiong. 2025. AI and Freelancers: Has the Inflection Point Arrived? In Proceedings of the 58th Hawaii International Conference on System Sciences.

Raviv, Shir. 2023. "When Do Citizens Resist The Use of Algorithmic Decision-making in Public Policy? Theory and Evidence." SSRN Electronic Journal .

Saccardo, Silvia, Aniela Pietrasz and Uri Gneezy. 2018. "On the Size of the Gender Difference in Competitiveness." Management Science 64(4):1541–1554.

Samek, Anya. 2019. "Gender Differences in Job Entry Decisions: A University-Wide Field Experiment." Management Science 65(7):3272–3281.

Schiff, Daniel S, Kaylyn Jackson Schiff and Robin Jacobson. 2024. The Politics of AI: Will Bipartisanship Last or Is Polarization Inevitable? In MPSA 2024. Chicago: .

Schmidgall, Samuel, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu and Emad Barsoum. 2025. "Agent Laboratory: Using LLM Agents as Research Assistants.". arXiv:2501.04227 [cs].

Simpson, E. H. 1951. "The Interpretation of Interaction in Contingency Tables." Journal of the Royal Statistical Society Series B: Statistical Methodology 13(2):238–241.

Smith, Ray A. 2024. "AI Is Starting to Threaten White-Collar Jobs. Few Industries Are Immune." The Wall Street Journal .

Trajtenberg, Manuel. 2019. AI as the Next GPT: A Political-Economy Perspective. In The Economics of Artificial Intelligence: An Agenda, ed. Ajay Agrawal, Joshua S. Gans and Avi Goldfarb. Chicago: University of Chicago Press pp. 175–186.

Wagenmakers, Eric-Jan. 2007. "A practical solution to the pervasive problems of p values." Psychonomic Bulletin & Review 14(5):779–804.

Appendix: For Online Publication

Contents

A Full Model Output

Table S1: Model results for Figure 3

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−134.010***	−234.902***	−101.230***	1.262***	0.645***	2.137***
	(14.548)	(18.990)	(17.993)	(0.122)	(0.095)	(0.117)
Constant	314.465***	368.981***	323.082***	2.982***	2.392***	1.909***
	(7.554)	(9.599)	(8.625)	(0.063)	(0.048)	(0.056)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.094	0.174	0.045	0.081	0.060	0.354
Panel B: First stage						
Treatment	0.666***	0.668***	0.631***	0.666***	0.668***	0.631***
	(0.022)	(0.022)	(0.023)	(0.022)	(0.022)	(0.023)
Constant	0.076***	0.056***	0.056***	0.076***	0.056***	0.056***
	(0.016)	(0.015)	(0.016)	(0.016)	(0.015)	(0.016)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.462	0.474	0.430	0.462	0.474	0.430
F Statistic (df = 1; 1039)	891.139***	935.435***	785.369***	891.139***	935.435***	785.369***

Note: Full results for Figure 3. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S2: Model results for Figure 4

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−139.413***	−230.298***	−95.276***	1.334***	0.459***	2.000***
	(18.965)	(24.290)	(23.208)	(0.158)	(0.122)	(0.152)
Frequent AI user	−10.035	18.297	5.948	−0.397*	−0.346**	−0.533***
	(20.373)	(25.797)	(23.495)	(0.170)	(0.130)	(0.153)
Used AI × Frequent AI user	17.484	−19.931	−17.493	0.006	0.623**	0.596*
	(32.306)	(42.510)	(40.234)	(0.270)	(0.214)	(0.263)
Constant	316.339***	365.287***	322.157***	3.070***	2.457***	2.015***
	(8.137)	(10.360)	(9.228)	(0.068)	(0.052)	(0.060)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.092	0.173	0.046	0.090	0.048	0.344
Panel B: First stage						
Treatment	0.614***	0.628***	0.587***	0.614***	0.628***	0.587***
	(0.026)	(0.025)	(0.026)	(0.026)	(0.025)	(0.026)
Frequent AI user	0.189***	0.174***	0.185***	0.189***	0.174***	0.185***
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Treatment × Frequent AI user	0.077	0.051	0.057	0.077	0.051	0.057
	(0.047)	(0.047)	(0.048)	(0.047)	(0.047)	(0.048)
Constant	0.032	0.015	0.012	0.032	0.015	0.012
	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)	(0.017)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.509	0.510	0.472	0.509	0.510	0.472
F Statistic (df = 3; 1037)	357.680***	359.590***	309.546***	357.680***	359.590***	309.546***

Note: Full results for Figure 4. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S3: Model results for Figure 6/7 - Panel A: Sex

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−141.754***	−245.248***	−121.829***	1.405***	0.659***	2.091***
	(17.851)	(22.665)	(21.717)	(0.151)	(0.113)	(0.141)
Female	−1.505	−20.563	−22.446	0.319*	0.018	0.079
	(15.940)	(20.438)	(18.289)	(0.134)	(0.102)	(0.119)
Used AI × Female	23.157	33.796	64.484	−0.441	−0.045	0.137
	(30.790)	(41.507)	(38.598)	(0.260)	(0.208)	(0.250)
Constant	315.145***	375.628***	330.498***	2.877***	2.387***	1.884***
	(9.300)	(11.720)	(10.522)	(0.078)	(0.059)	(0.068)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.095	0.174	0.050	0.075	0.059	0.356
Panel B: First stage						
Treatment	0.659***	0.681***	0.633***	0.659***	0.681***	0.633***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Female	−0.034	−0.021	−0.021	−0.034	−0.021	−0.021
	(0.033)	(0.033)	(0.034)	(0.033)	(0.033)	(0.034)
Treatment × Female	0.021	−0.037	−0.008	0.021	−0.037	−0.008
	(0.048)	(0.047)	(0.048)	(0.048)	(0.047)	(0.048)
Constant	0.087***	0.063***	0.063***	0.087***	0.063***	0.063***
	(0.019)	(0.018)	(0.019)	(0.019)	(0.018)	(0.019)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.462	0.476	0.431	0.462	0.476	0.431
F Statistic (df = 3; 1037)	297.216***	313.427***	261.953***	297.216***	313.427***	261.953***

Note: Full results for Figure 6/7 - Panel A: Sex. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S4: Model results for Figure 6/7 - Panel B: Age

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	-153.773^{***}	-268.751^{***}	-103.761^{***}	1.218^{***}	0.653^{***}	2.267^{***}
	(21.911)	(28.516)	(28.341)	(0.184)	(0.142)	(0.184)
Under 40	-25.488	-44.801^{*}	-21.706	-0.104	0.138	0.129
	(15.066)	(19.134)	(17.125)	(0.127)	(0.096)	(0.111)
Used AI × Under 40	36.034	61.717	3.525	0.079	-0.010	-0.229
	(29.264)	(38.146)	(36.502)	(0.246)	(0.190)	(0.236)
Constant	327.306^{***}	391.692^{***}	334.533^{***}	3.036^{***}	2.319^{***}	1.844^{***}
	(10.396)	(13.272)	(11.814)	(0.087)	(0.066)	(0.077)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.093	0.175	0.048	0.080	0.064	0.356
Panel B: First stage						
Treatment	0.639^{***}	0.643^{***}	0.578^{***}	0.639^{***}	0.643^{***}	0.578^{***}
	(0.032)	(0.031)	(0.032)	(0.032)	(0.031)	(0.032)
Under 40	0.077^{*}	0.068^{*}	0.076^{*}	0.077^{*}	0.068^{*}	0.076^{*}
	(0.031)	(0.030)	(0.031)	(0.031)	(0.030)	(0.031)
Treatment × Under 40	0.054	0.052	0.106^{*}	0.054	0.052	0.106^{*}
	(0.044)	(0.043)	(0.044)	(0.044)	(0.043)	(0.044)
Constant	0.036	0.020	0.016	0.036	0.020	0.016
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.474	0.484	0.451	0.474	0.484	0.451
F Statistic (df = 3; 1037)	310.992^{***}	323.915^{***}	283.954^{***}	310.992^{***}	323.915^{***}	283.954^{***}

Note: Full results for Figures 6/7 - Panel B: Age. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S5: Model results for Figure 6/7 - Panel C: Education

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	-124.111^{***}	-191.893^{***}	-91.311^{*}	1.094^{***}	0.626^{**}	2.248^{***}
	(32.341)	(41.054)	(42.779)	(0.270)	(0.205)	(0.275)
Degree	10.911	15.960	10.632	0.112	0.129	0.189
	(18.091)	(22.945)	(20.901)	(0.151)	(0.115)	(0.134)
Used AI × Degree	−11.028	−55.758	−14.318	0.215	0.037	−0.122
	(36.303)	(46.408)	(47.327)	(0.303)	(0.232)	(0.305)
Constant	305.677^{***}	357.486^{***}	316.082^{***}	2.896^{***}	2.291^{***}	1.758^{***}
	(15.885)	(20.143)	(18.414)	(0.133)	(0.101)	(0.118)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.094	0.176	0.044	0.085	0.063	0.358
Panel B: First stage						
Treatment	0.665^{***}	0.684^{***}	0.588^{***}	0.665^{***}	0.684^{***}	0.588^{***}
	(0.049)	(0.048)	(0.050)	(0.049)	(0.048)	(0.050)
Degree	0.046	0.046	0.037	0.046	0.046	0.037
	(0.038)	(0.038)	(0.039)	(0.038)	(0.038)	(0.039)
Treatment × Degree	0.002	−0.020	0.048	0.002	−0.020	0.048
	(0.055)	(0.054)	(0.056)	(0.055)	(0.054)	(0.056)
Constant	0.037	0.018	0.028	0.037	0.018	0.028
	(0.034)	(0.034)	(0.035)	(0.034)	(0.034)	(0.035)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.466	0.475	0.427	0.466	0.475	0.427
F Statistic (df = 3; 1023)	297.292^{***}	307.926^{***}	254.539^{***}	297.292^{***}	307.926^{***}	254.539^{***}

Note: Full results for Figures 6/7 - Panel C: Education. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S6: Model results for Figure 6/7 - Panel D: Occupation

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−131.182***	−205.098***	−109.294***	1.324***	0.604***	2.161***
	(26.415)	(34.547)	(30.318)	(0.222)	(0.172)	(0.196)
Professional	13.961	18.813	−7.477	0.138	0.155	0.203
	(15.872)	(20.080)	(17.925)	(0.133)	(0.100)	(0.116)
Used AI × Professional	−4.803	−43.670	12.397	−0.095	0.050	−0.048
	(31.667)	(41.413)	(37.703)	(0.266)	(0.206)	(0.243)
Constant	305.279***	356.749***	327.921***	2.891***	2.290***	1.775***
	(12.828)	(16.127)	(14.277)	(0.108)	(0.080)	(0.092)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.095	0.173	0.044	0.083	0.065	0.358
Panel B: First stage						
Treatment	0.643***	0.645***	0.657***	0.643***	0.645***	0.657***
	(0.039)	(0.038)	(0.039)	(0.039)	(0.038)	(0.039)
Professional	0.016	0.026	0.035	0.016	0.026	0.035
	(0.033)	(0.032)	(0.033)	(0.033)	(0.032)	(0.033)
Treatment × Professional	0.033	0.033	−0.040	0.033	0.033	−0.040
	(0.048)	(0.047)	(0.048)	(0.048)	(0.047)	(0.048)
Constant	0.066*	0.038	0.033	0.066*	0.038	0.033
	(0.027)	(0.026)	(0.027)	(0.027)	(0.026)	(0.027)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.463	0.476	0.431	0.463	0.476	0.431
F Statistic (df = 3; 1037)	297.833***	313.582***	261.939***	297.833***	313.582***	261.939***

Note: Full results for Figures 6/7 - Panel D: Occupation. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

B Additional Details About the Survey

Sampling

To ensure the feasibility of the study, YouGov asked 34,211 people between 14th April and 4th May 2023 whether they have a ChatGPT account. The study sample is recruited from the 5,350 respondents who stated that they have an account (free or paid). Below we list the feasibility question and the distribution of answers.

- ChatGPT is an AI-based computer program that can generate human-like text. Before taking this survey, had you ever used ChatGPT?
  - Yes, and I have a paid account (n=865)
  - Yes, and I have a free account (n=4485)
  - Yes, but I used someone else's account (n=774)
  - No, I have never used ChatGPT (n=28087)

While we acknowledge that this introduces some selection issues, we employ quotas for age, gender, income, and region to ensure a broadly representative sample. Given the user pool at the time, we did not manage to obtain a fully representative sample. As Table S7 shows, the sample is more male and more educated than our target population, the UK working age population. However, Table S8 shows good balance between the treatment and control groups and Table S22 shows that the results are robust to controlling for demographic characteristics. Moreover, any residual bias is likely to be conservative, as participants in the control group already have an account and could use ChatGPT with little extra effort.

Table S7: Sample characteristics

Characteristic	Count/Mean (%/SD) Treatment	Count/Mean (%/SD) Control
Sex: Female	165 (32.7%)	170 (31.7%)
Sex: Male	339 (67.3%)	367 (68.3%)
Uni: Yes	393 (79.1%)	421 (79.4%)
Uni: No	104 (20.9%)	109 (20.6%)
Professional: Yes	346 (68.7%)	355 (66.1%)
Professional: No	158 (31.4%)	182 (33.9%)
Age	40.5 (11.2)	39.8 (11.3)
HH Income	10.9 (3.2)	10.6 (3.4)
Skills Average	1.9 (0.5)	2.0 (0.6)

We pre-tested a version of the survey which included a control group in which we did not prime participants about AI at all. Since answer quality in this group was generally lower and a substantial share of respondents used ChatGPT, we did not include this control group in the final version of the survey.

Table S8: Balance Tests

Variable name	Type	Std. Mean Dif.
age	C	0.0616
sex: Male	B	-0.0108
degree	B	-0.0036
degree: NA	B	0.0009
prof_man	B	0.0254
hh_income	C	0.1017*
hh_income: NA	B	-0.0116
skills_avg	C	-0.0460
skills_avg: NA	B	0.0010
education: GCSEs / O-Levels or none	B	0.0046
education: A-Levels or equivalent	B	-0.0010
education: Undergraduate	B	-0.0216
education: Postgraduate	B	0.0180

Note: B = binary variable; C = continuous variable. Sample sizes: Control: 537; Treatment: 504.

Inter-Coder Reliability

Two of the authors each coded a random subset of 100 answers to assess the reliability of the ChatGPT grading. Table S9 shows that average inter-coder agreement between ChatGPT and human coders was similar to human – human inter-coder agreement, suggesting that a well-designed AI-assisted grading scheme can lead to substantial time savings without loss of quality. This illustrates the potential of LLMs to drastically increase the efficiency of coding text data for quantitative analysis.

Table S9: Inter-coder reliability between ChatGPT and human coders

	GPT-4 & R1			**GPT-4 & R2**			**R1 & R2**		
Task	**O**	T	C	**O**	T	C	**O**	T	C
Email	**0.702**	0.764	0.545	**0.662**	0.750	0.488	**0.851**	0.817	0.837
	(5.72)	(7.47)	(3.39)	**(4.92)**	(6.69)	(2.91)	**(12.4)**	(9.93)	(11.2)
Assessment	**0.720**	0.670	0.784	**0.770**	0.741	0.788	**0.905**	0.893	0.924
	(6.15)	(5.07)	(8.25)	**(7.69)**	(6.72)	(8.42)	**(20.1)**	(17.6)	(25.2)
Comprehension	**0.940**	0.958	0.883	**0.888**	0.835	0.878	**0.874**	0.846	0.849
	(32.3)	(46.2)	(16.1)	**(16.9)**	(11.1)	(15.4)	**(14.9)**	(12.0)	(12.3)

Note: Intra-class correlations (F-statistics in parentheses) calculated based on a random sample of 100 responses that were also graded by two of the authors (R1 & R2). **O:** overall sample (N = 100); T: treatment group (N = 51); C: control group (N = 49).

Table S10: Comparison of means in the inter-coder subsample

	Email			**Assessment**			**Comprehension**		
Grader	C	T	Δ	C	T	Δ	C	T	Δ
GPT-4	3.3	4.1	0.8	2.4	2.9	0.5	2.1	3.5	1.4
	(1.3)	(1.3)		(1.0)	(1.3)		(1.1)	(1.4)	
R1	3.1	4.0	0.9	2.8	3.1	0.3	2.4	3.5	1.1
	(1.0)	(1.2)		(1.2)	(1.4)		(1.2)	(1.5)	
R2	3.4	4.3	0.9	2.7	3.3	0.6	2.3	3.8	1.5
	(1.2)	(1.1)		(1.1)	(1.3)		(1.1)	(1.3)	

Note: Means (standard deviations in parentheses) calculated based on a random sample of 100 responses that were also graded by two of the authors (R1 & R2). C: control group (N = 49); T: treatment group (N = 51); **Δ:** mean difference between control and treatment group.

Deviations from the Pre-Analysis Plan

Some changes were made compared to the procedures outlined in the pre-analysis plan. Here we detail these changes and the reason why they were necessary.

- **Small changes to the grading scheme:** Some very minor changes to the grading scheme were necessary to improve the workflow. The full grading scheme as it was used in the final analysis is attached in Appendix E.
- **Instrumental variable estimation:** We specified that we would estimate OLS or linear probability models to calculate the intention-to-treat effect (ITT). In order to estimate the local average treatment effect (LATE) instead of the ITT, we present 2SLS estimations using treatment assignment as an instrumental variable for self-reported compliance with the treatment instructions. Models estimated with OLS are included in the supplementary materials.
- **Reformulation of the hypotheses:** Reflecting the shift from ITT to LATE estimation, we slightly reformulated the hypotheses to reflect the analysis of AI use, rather than AI exposure.
- **Verification of compliance:** In addition to manually inspecting answers to identify misclassifications of ChatGPT use, we verified that compliers in the control group do not perform significantly different from non-compliers in the treatment group. As described in Appendix D, this analysis confirms that respondents' self-reports are largely accurate.

Descriptive Statistics

Before the first task, we asked respondents to assess their own skills on a number of dimensions: understanding written information, writing informative texts, communicating with others, performing under pressure, using new technologies, and English proficiency. We find no meaningful differences between the treatment and control group on any of the measures. Respondents overwhelmingly described themselves as "somewhat skilled" in all five dimensions, as we show in Figure S1. Furthermore, the vast majority of respondents are native English speakers, as can be seen in Figure S2. After each task, we asked respondents how frequently they perform tasks similar to the one they just did in their daily life. As Figure S3 shows, familiarity with the tasks is very similar in both groups. The levels of familiarity correspond to task complexity, with people being most likely to be familiar with the email task and least likely with the comprehension task. This, alongside the balance tests on demographic characteristics reported in Table S8, provides reassurance that the treatment effects detailed below do not result from different characteristics of the samples.

Table S11: Performance on the tasks by treatment

		Treatment	
Task	**Variable**	**Control**	**Treatment**
Email	Median score	3	5
	Mean score (SD)	3.1 (1.2)	3.9 (1.4)
	Median time	263.2	182.4
Assessment	Median score	2	3
	Mean score (SD)	2.4 (0.9)	2.9 (1.2)
	Median time	287.3	132.4
Comprehension	Median score	1.7	3.7
	Mean score (SD)	2.0 (1.0)	3.4 (1.6)
	Median time	272.3	212.6

One potential objection to our results is that they might be influenced by the type of device respondents used to complete the survey. In particular, people using a mobile device might find it more onerous to access ChatGPT and read the texts that make up the tasks, leading to lower ChatGPT uptake in the treatment group and lower response quality. To address this concern, we recorded the operating system of the device used to complete the survey, based on which we created a dummy for respondents who used a mobile device (approximately 70% of the sample). In Table S12 we show that mobile users are evenly distributed across treatment groups. Self-reported AI uptake indeed tends to be slightly lower among users of mobile devices, by about 10 percentage points in the control group and by up to 7 percentage points in the treatment group. Nevertheless, performance, both in terms of grades and time spent on the tasks, is very similar for respondents who used a mobile device and those who used a laptop or desktop computer, with the possible exception of the comprehension task. Consequently, we find virtually identical treatment effects when we restrict the sample to mobile phone users only and we conclude that our findings are not driven by differences in the type of devices used.

Table S12: Compliance by treatment group and device type

		Control group		Treatment group	
Task		Mobile device	Other device	Mobile device	Other device
Email	AI: No	367 (95%)	129 (85%)	88 (25%)	42 (27%)
	AI: Yes	19 (5%)	22 (15%)	258 (75%)	116 (73%)
	Mean score (SD)	3.1 (1.2)	3.1 (1.2)	3.9 (1.4)	3.9 (1.4)
Assessment	AI: No	375 (97%)	132 (87%)	103 (30%)	36 (23%)
	AI: Yes	11 (3%)	19 (13%)	243 (70%)	122 (77%)
	Mean score (SD)	2.4 (0.9)	2.4 (0.8)	2.8 (1.2)	2.9 (1.2)
Comprehension	AI: No	377 (98%)	130 (86%)	115 (33%)	43 (27%)
	AI: Yes	9 (2%)	21 (14%)	231 (67%)	115 (73%)
	Mean score (SD)	1.9 (1.0)	2.3 (1.1)	3.3 (1.6)	3.6 (1.5)

Note: Cross-classification of treatment group by type of device used. The share of mobile device users is slightly higher in the control group, as is compliance with the instructions. Within-treatment group performance is not affected by type of device used, except in the comprehension task.

Figure S1: Skill self-assessments

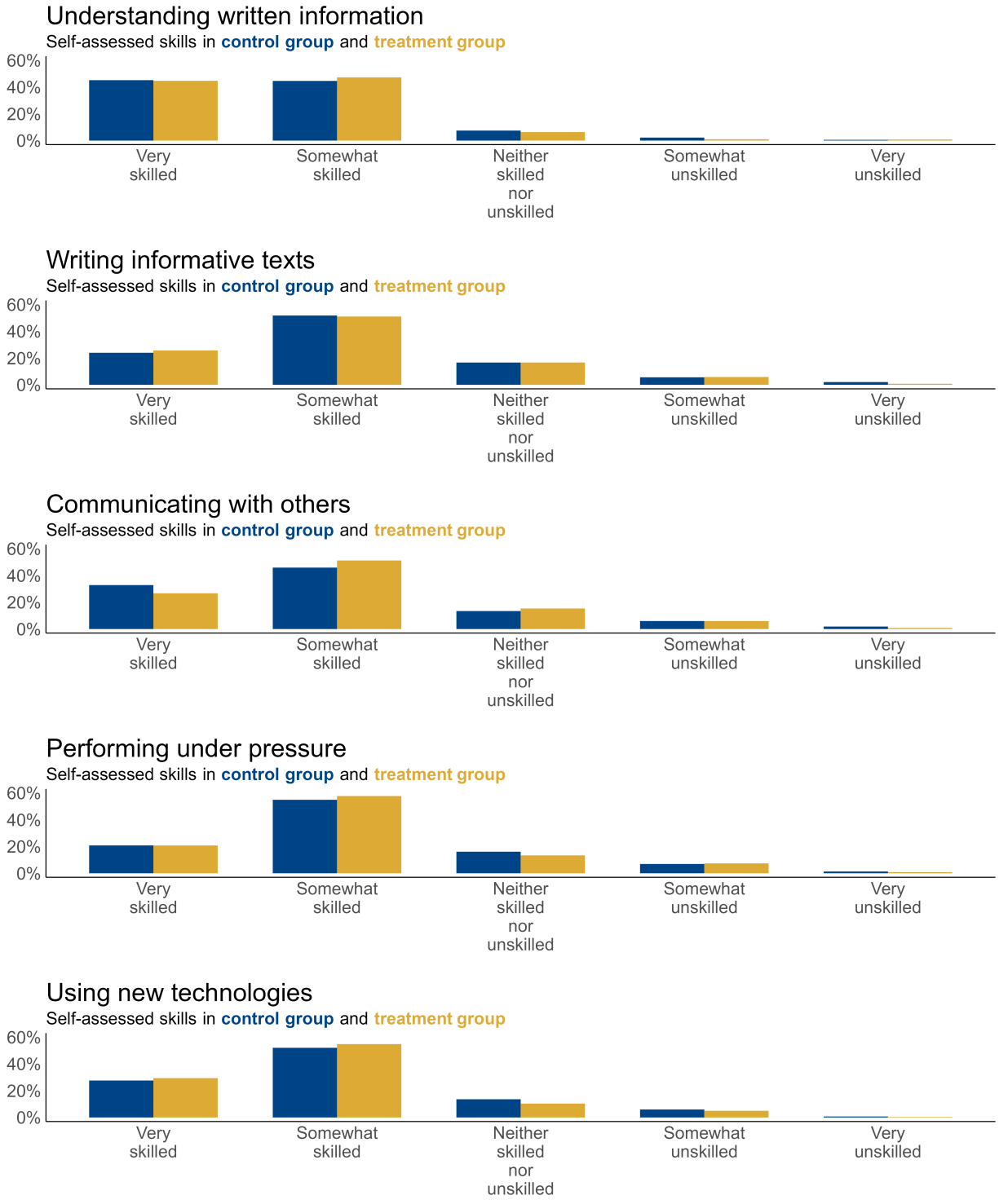


Figure S2: English proficiency

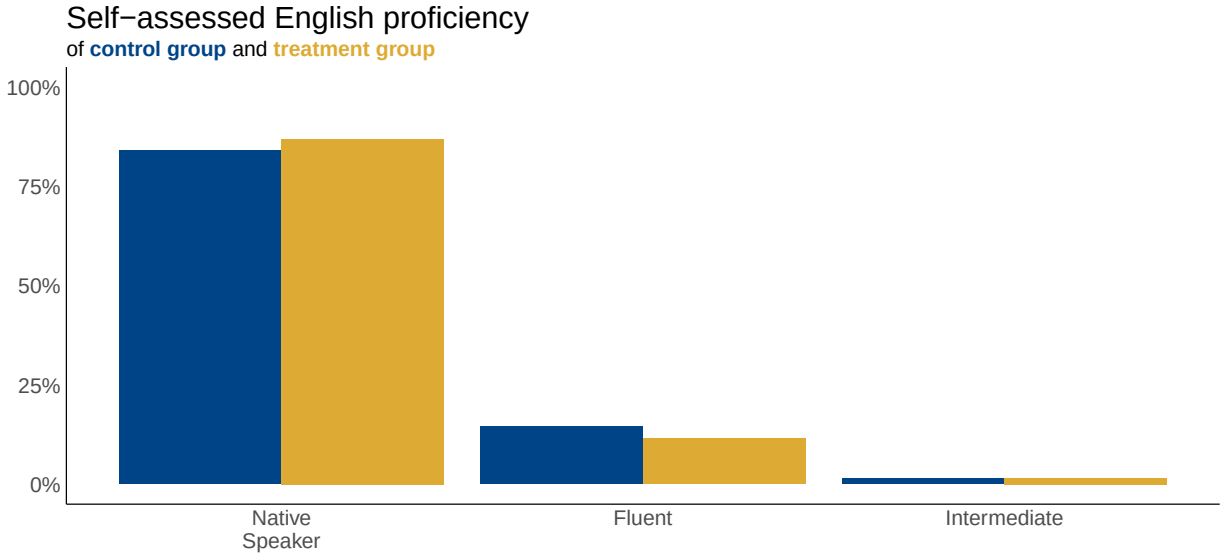
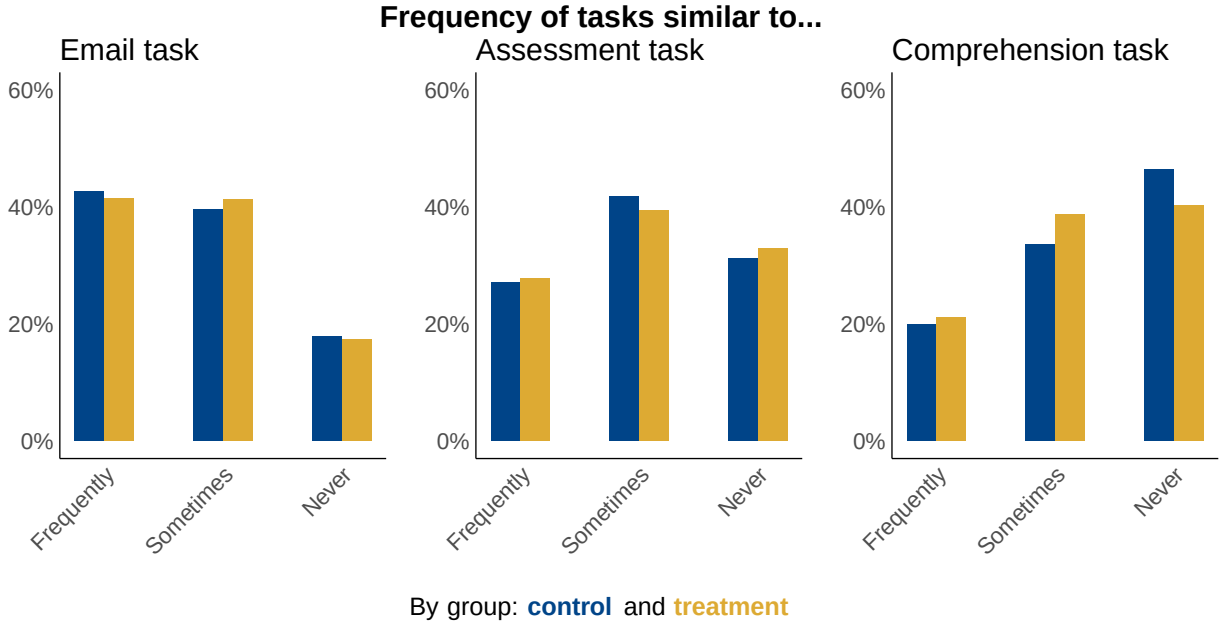
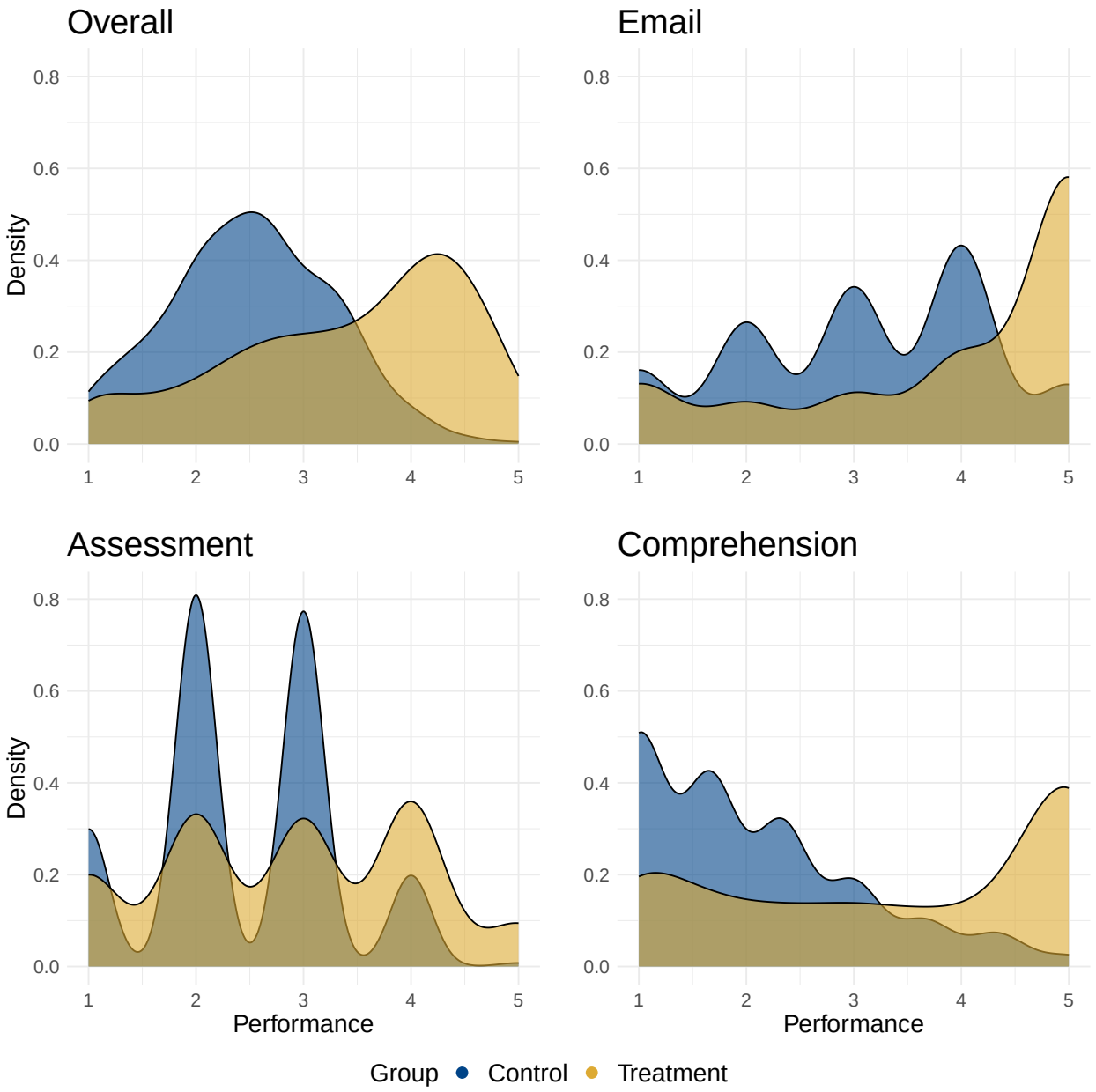


Figure S3: Familiarity of respondents with the tasks



Note: The figure shows whether respondents perform the task frequently (once a week or more often), sometimes, or never.

Figure S4: Density plots of performance by treatment group



C Power Analysis and Bayes Factors

Due to the partly counterintuitive findings of this study, it is imperative to be transparent about the statistical power of the underlying evidence. Funding constraints prevented us from conducting our survey experiment with a larger sample. Nevertheless, evidence from power analyses and Bayes Factor analyses indicates that a high degree of confidence in the key findings – notably the absence of an interaction between the treatment and sex, education, or occupational group – is warranted. In a nutshell, simulated power curves show that for a realistic range of parameters, we are able to detect small- to medium-sized interaction effects. A Bayes factor analysis finds strong support for the null results of our conventional statistical analysis. We now discuss these analyses in more detail.

To calculate our statistical power for subgroup analyses, we performed power calculations on simulated datasets with a realistic range of parameters based on our observed data. We apply the conventional threshold of 80 percent power and a significance level of 0.05. We express effect sizes in terms of Cohen's delta (effect size divided by the standard deviation) for easier interpretability. The key factor that influences how small an interaction effect can be and still be detectable, is the balance of group sizes in the moderator variable: the more unequal the size of the moderator groups, the larger the required sample size. While the probability of being assigned to the treatment group is always 0.5, the we run simulations for three different splits of the moderator variable: 50/50, 65/35, and 80/20. Other parameters are based on the values observed in the sample.[21] We perform 1000 simulations for each of the 216 combinations of parameters.

Figure S5 shows this analysis. The bands in the figure circumscribe the minimum and maximum of the 9 power estimates for each given size of the interaction effect and group split. Conventionally, a Cohen's d of 0.2 is considered a small effect, while 0.5 and 0.8 are considered a medium and large effect, respectively. This means that we have sufficient power to detect small- to medium-sized interaction effects where the moderator variable is distributed approximately in a 35/65 split (as is the case for sex and occupational group) and medium-sized effects when the distribution of the moderator is more skewed (as is the case for education). However, the interaction effects that we find in our analyses are mostly very small, with a Cohen's d well below 0.2.[22] Figure S6 shows that even with 4000 participants, we would not find these effects to be statistically significant at the 5 percent level and with 80 percent power. It is therefore safe to conclude that factors which predict frequent AI use, such as sex, educational attainment, and broad occupational categories, have at best a negligible impact on how effectively people use AI in common text-based tasks.

[21] The intercept and standard deviation are based on the values for each task. Figure S5 takes the values from the email task, but substantively identical results are available for the assessment and comprehension tasks. N = 1041 is the number of observations in our sample. The Cohen's d values for the main effects are 0.5, 0.8, and 1.0 for the treatment and 0.1, 0.2, and 0.35 for the moderator.

[22] To calculate Cohen's d, divide the interaction coefficient by the standard deviation, which is 1.37, 1.05, and 1.47 in the email, assessment, and comprehension tasks, respectively.

Figure S5: Power Analysis for Realistic Parameters Given Sample Characteristics

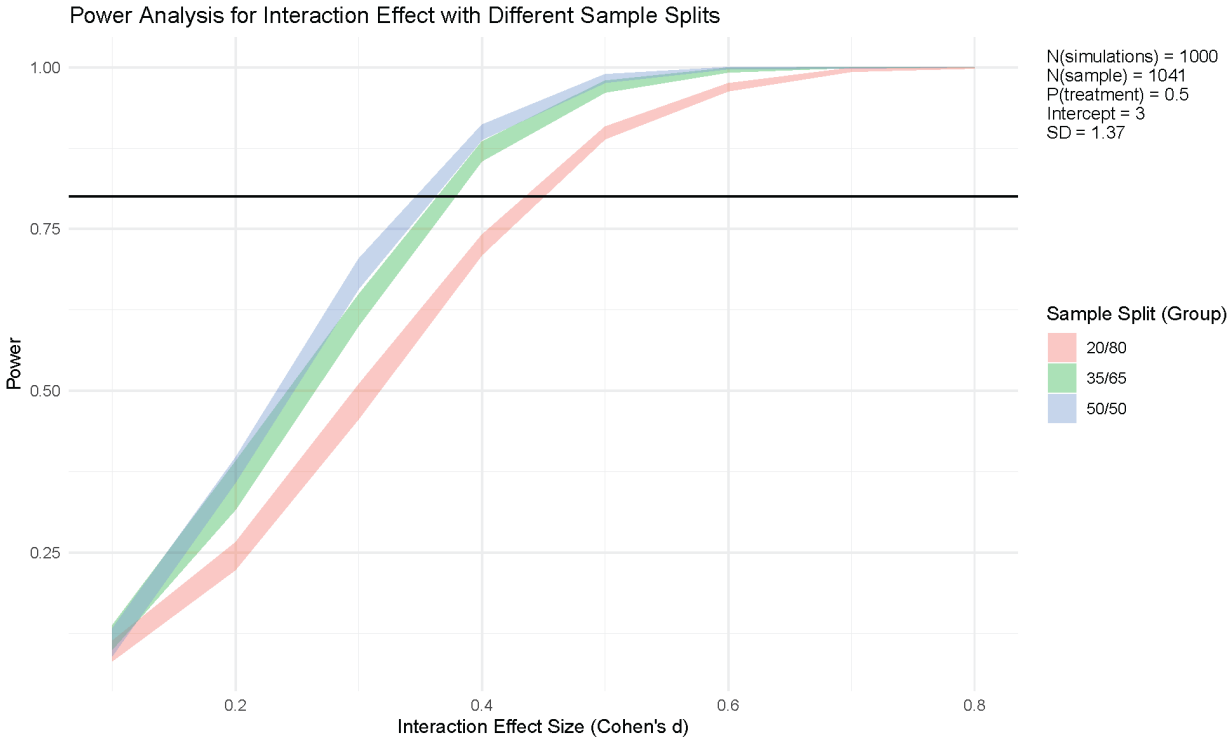
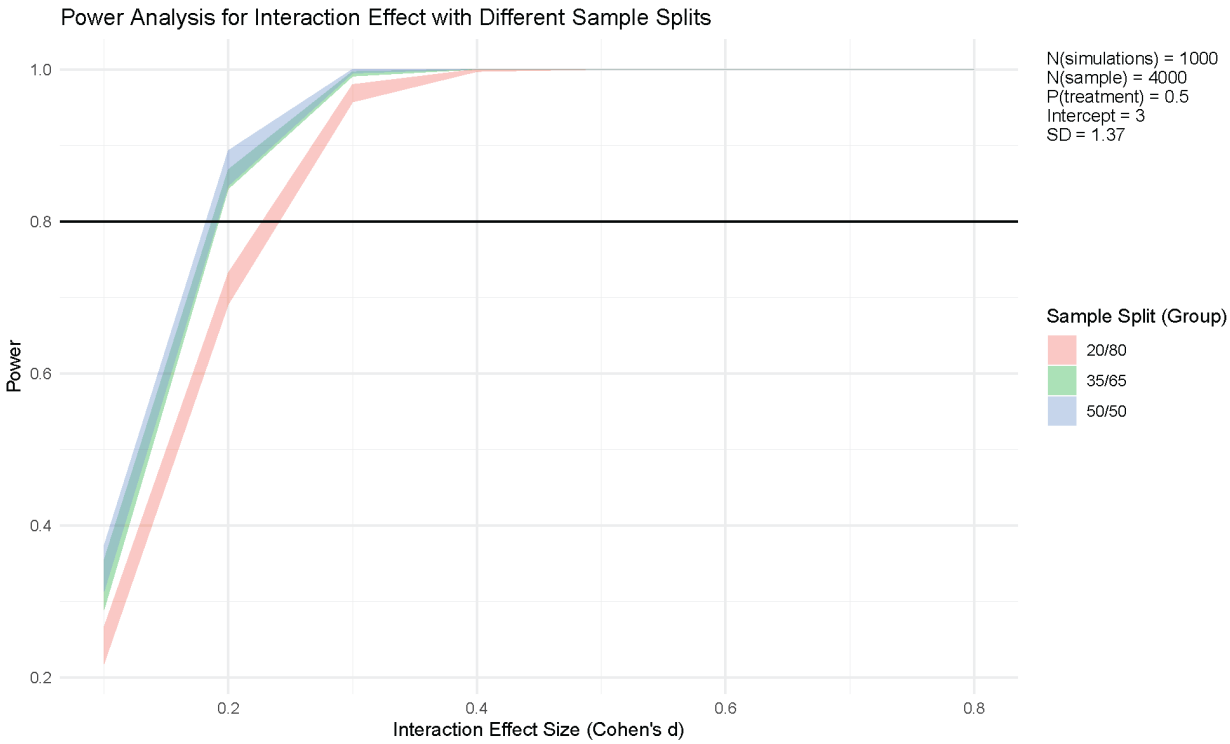


Figure S6: Power Analysis for Hypothetical Larger Sample



Bayes factor analysis offers complementary evidence and strengthens our findings. The Bayes factor fulfils a function similar to the ubiquitous p-values, but has several advantages (Jarosz and Wiley, 2014; Lin and Yin, 2015). Most notably, the Bayesian approach compares the amount of evidence for different hypotheses, whereas the frequentist approach to null hypothesis significance testing "considers only the extremeness of the data under the null hypothesis, with no consideration of the alternative hypothesis" ((Jarosz and Wiley, 2014, 3); Wagenmakers (2007)). Thus, "the Bayes factor does not render a dichotomous (reject or not reject H_0) decision", but instead quantifies the support in the data for the various hypotheses under consideration (Hoijtink et al., 2019, 20). This is especially helpful in the context of our null findings. Rather than reject the null hypothesis (or fail to do so), we can compare the evidence for a null interaction effect to the evidence for a positive or negative interaction effect.

We use the *bain* package in R to compute Bayes factors and posterior probabilities based on ITT estimates for the following three hypotheses for each of the analyses: $moderator \times treatment < 0; moderator \times treatment = 0; moderator \times treatment > 0$. That is, we test hypotheses that cover the entire parameter space. In Table S13, we report two measures of the Bayes factor alongside the posterior model probability. A higher Bayes factor and a higher posterior probability indicate greater support for the hypothesis relative to the other hypotheses. The column "BF (unconstrained)" captures how the hypothesis compares to an unconstrained model, whereas the column "BF (complement)" reflects how it fares against the other hypotheses specified in the analysis. The posterior model probability provides an overall measure of relative support for the hypotheses and is the main basis for our conclusions.

The results bolster the findings of the frequentist analysis, as the posterior model probabilities show that the absence of an interaction between the respective moderators and the treatment is by far the likeliest scenario in all of our models.[23] For reasons of space, in Tables S13 - S15 we present only the results for the analyses of the task scores. Tables for the analyses with completion time show the same qualitative pattern and are available upon request. The posterior probabilities of the hypothesis that there is no interaction with the treatment are above 0.9 in most cases, providing additional evidence that the null relationships we find are not merely due to a limited sample size. In this context, it is instructive to also consider the Bayes factors themselves. In one case (sex in the email task), the Bayes factor complement appears to contradict the posterior probability and the unconstrained Bayes factor, as well as the significance test in the frequentist analysis. In this instance, the totality of the evidence can be interpreted to indicate that while a null interaction is most likely to be true, conditional on there being a relationship, it is of the direction that the Bayes factor complement indicates. In other words, the Bayes factor complement cautions against

[23] In the models without an interaction, the Bayesian analysis shows definitive evidence for a positive effect of the treatment on all outcomes, in line with the main analyses presented in the paper.

Table S13: Bayes Factors for Interaction Effects (Performance in Email Task)

Hypothesis: Treatment *	BF (unconstrained)	BF (complement)	Posterior Probability
Age under 40 < 0	0.46	0.30	0.02
Age under 40 $= 0$	24.43	24.43	0.92
Age under 40 > 0	1.54	3.39	0.06
Female < 0	1.88	16.05	0.16
Female $= 0$	9.47	9.47	0.83
Female > 0	0.12	0.06	0.01
Degree < 0	0.47	0.30	0.02
Degree $= 0$	24.59	24.59	0.92
Degree > 0	1.53	3.29	0.06
Professional < 0	1.10	1.22	0.03
Professional $= 0$	32.02	32.02	0.94
Professional > 0	0.90	0.82	0.03

Table S14: Bayes Factors for Interaction Effects (Performance in Assessment Task)

Hypothesis: Treatment *	BF (unconstrained)	BF (complement)	Posterior Probability
Age under 40 < 0	0.83	0.71	0.02
Age under 40 $= 0$	31.55	31.55	0.94
Age under 40 > 0	1.17	1.40	0.03
Female < 0	1.30	1.87	0.04
Female $= 0$	29.90	29.90	0.94
Female > 0	0.70	0.53	0.02
Degree < 0	0.94	0.89	0.03
Degree $= 0$	31.95	31.95	0.94
Degree > 0	1.06	1.13	0.03
Professional < 0	0.69	0.53	0.02
Professional $= 0$	29.83	29.83	0.94
Professional > 0	1.31	1.89	0.04

completely dismissing the possibility of an interaction effect based on the posterior model probabilities. Overall, the Bayes factors and posterior probabilities show that where we find null relationships in the frequentist analysis, we can be confident that there really is no relationship.

References in this Section

- Hoijtink H, Mulder J, Van Lissa C, Gu X. A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, October 2019; 24(5):539–556.
- Jarosz AF, Wiley J. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, November 2014; 7(1).
- Lin R, Yin G. Bayes factor and posterior probability: Complementary statistical evidence to p-value. *Contemporary Clinical Trials*, September 2015; 44:33–35.
- Wagenmakers EJ. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin &*

Table S15: Bayes Factors for Interaction Effects (Performance in Comprehension Task)

Hypothesis: Treatment *	**BF (unconstrained)**	**BF (complement)**	**Posterior Probability**
Age under 40 < 0	0.60	0.43	0.02
Age under 40 $= 0$	28.22	28.22	0.93
Age under 40 > 0	1.40	2.31	0.05
Female < 0	0.69	0.53	0.02
Female $= 0$	29.84	29.84	0.94
Female > 0	1.31	1.89	0.04
Degree < 0	0.88	0.79	0.03
Degree $= 0$	31.70	31.70	0.94
Degree > 0	1.12	1.27	0.03
Professional < 0	1.49	2.95	0.05
Professional $= 0$	25.88	25.88	0.93
Professional > 0	0.51	0.34	0.02

Review, October 2007; 14(5):779–804.

D Compliance with Treatment Instructions

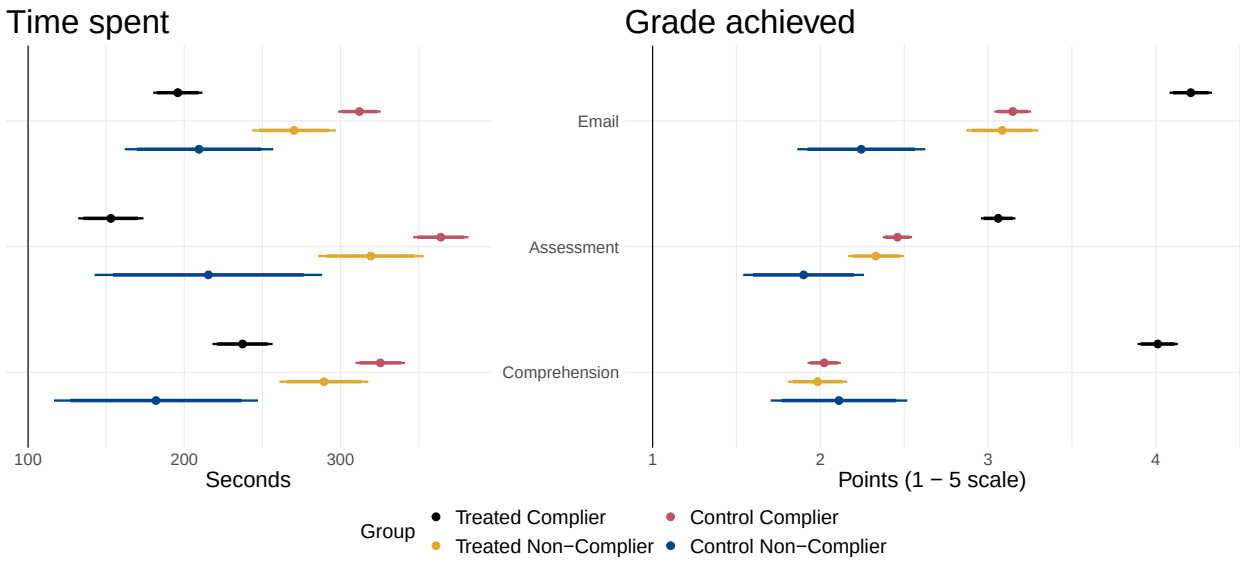
Self-Reported Compliance and Treatment Effects

Respondents in the treatment group may refrain from using ChatGPT, while respondents in the control group may defy instructions and use AI even when they should not. Designing the survey experiment to maximize compliance was therefore paramount. We structured our financial incentives to increase compliance with the instructions to use or not use ChatGPT. We highlighted the prospect of additional rewards in case of the treatment group and the threat of being excluded from the survey for non-compliance in case of the control group. In addition to visually inspecting the answers, we asked respondents after each task to self-report which, if any, online tools they used. People generally report high compliance with the instructions: well over 90% in the control group state for each task that they did not use AI, and between two thirds and three quarters of respondents in the treatment group report that they did (see Table S12). We call these respondents compliers. In Figure S7 we plot the predicted values of time spent and grade achieved by treatment and compliance status.

The overall increase in performance in the treatment group is entirely driven by the compliers. Non-compliers in the treatment group perform very similar to compliers in the control group, implying that self-reports of AI use are generally truthful. Based on the reasonable assumption that people in the treatment group have no incentive to underreport their use of AI, we can compare compliers in the control group and non-compliers in the treatment group. If there was widespread underreporting of AI use in the control group, we would expect compliers in the control group to exhibit substantially higher performance than non-compliers in the treatment group. The low performance of the few non-compliers in the control group (between 30 and 41 individuals) is likely due to a number of inattentive respondents who exerted low effort on the tasks and selected random answers to the self-report question. Similarly, non-compliers in the treatment group spend less time and have slightly lower scores than control-compliers, indicating that they are the less motivated participants.

Since we do not have full control over compliance in either the treatment or control group, regressing the productivity measures on the treatment dummy only yields the intention-to-treat (ITT) effect. Moreover, selection into compliance is not random. For example, compliance in the control group increases with age, while it decreases with age in the treatment group (see Table S16). Thus, even though our diagnostic measures indicate high compliance on average and any residual bias is likely to be conservative, as undetected AI use in the control group and non-use in the treatment group are both likely to narrow performance gaps, we estimate local average treatment effects (LATE) in the main analyses and report ITT estimates in Tables S19 - S22. ITT effects can be interpreted as a lower-bound estimate of the true effect of AI use. Nevertheless,

Figure S7: Predicted values by treatment and compliance status



Note: The figure shows predicted values calculated from linear probability models interacting the treatment dummy with a complier dummy, with 90% and 95% confidence intervals (thick and thin lines). N = 1,041. Full model output in Table S19.

the above analysis illustrates, firstly, that self-reports of AI use are generally reliable, and secondly, that the estimated treatment effects are not due to some confounding attributes of the treatment group. This further strengthens the evidence that ChatGPT allows people to perform standard work tasks such as editing sensitive emails, appraising large quantities of written information, and extracting information from complicated texts both better and faster.

Turnitin Estimates of AI Use

We additionally used Turnitin's AI detection tool to verify whether respondents in the control group complied with the survey instructions to refrain from using AI chatbots to assist with their responses.[24] To do this, we uploaded a random selection of responses from the treatment and control groups for each of the tasks into Turnitin and compared AI detection scores from the two groups. We repeated this exercise six times with a different set of randomly selected responses. We used a random sample of responses rather than the full sample to comply with Turnitin's word limits for AI detection. In accordance with the maximum word limit, we randomly selected 40 answers for the email task, 100 answers for the assessment task, and 35 answers for the comprehension task.[25] Scores range from 0-100 and estimate the percentage of the document that

[24] Turnitin is a commercial plagiarism detection software that also offers AI detection tools.

[25] At the same time, individual answers were generally too short to run the detection algorithm on them, preventing us from identifying AI use in individual answers. The random selection of answers is replicable with our code, but the AI scores may not be as the technology evolves and the detection algorithm is updated by Turnitin. The reported scores are based on an analysis that we conducted on 9 January 2024.

Table S16: Compliance by treatment group and age

		Control group		Treatment group	
Task	Age	Complier	Non-complier	Complier	Non-complier
Email	15 - 35	186 (88%)	26 (12%)	159 (84%)	31 (16%)
	36 - 50	219 (95%)	12 (5%)	163 (75%)	53 (25%)
	51+	91 (97%)	3 (3%)	52 (53%)	46 (47%)
Assessment	15 - 35	194 (92%)	18 (8%)	153 (81%)	37 (19%)
	36 - 50	221 (96%)	10 (4%)	154 (71%)	62 (29%)
	51+	92 (98%)	2 (2%)	58 (59%)	40 (41%)
Comprehension	15 - 35	194 (92%)	18 (8%)	153 (81%)	37 (19%)
	36 - 50	220 (95%)	11 (5%)	144 (67%)	72 (33%)
	51+	93 (99%)	1 (1%)	49 (50%)	49 (50%)

Note: Cross-classification of age by treatment group and compliance status. Compliance increases with age in the control group and decreases with age in the treatment group.

is AI generated. As expected, the treatment group consistently received higher AI detection scores than the control group for all three tasks, as outlined in Table S17. AI scores are high for the email task for both groups because the original text was generated in part using ChatGPT — nonetheless, there is still a noticeable difference in the mean scores between treatment and control groups on this task. In the control group, over 600 answers to the assessment task and 210 answers to the comprehension task, the algorithm found no evidence of AI use.

Table S17: Comparison of Turnitin AI detection scores by treatment

Sample	**Email** (n=40)		**Assessment** (n=100)		**Comprehension** (n=35)	
	Treat.	Control	Treat.	Control	Treat.	Control
Sample 1	97	94	37	0	13	0
Sample 2	99	83	37	0	19	0
Sample 3	98	88	8	0	4	0
Sample 4	95	90	24	0	18	0
Sample 5	98	84	21	0	14	0
Sample 6	100	81	24	0	13	0
Mean	**97.8**	**86.7**	**25.2**	**0.0**	**13.5**	**0.0**

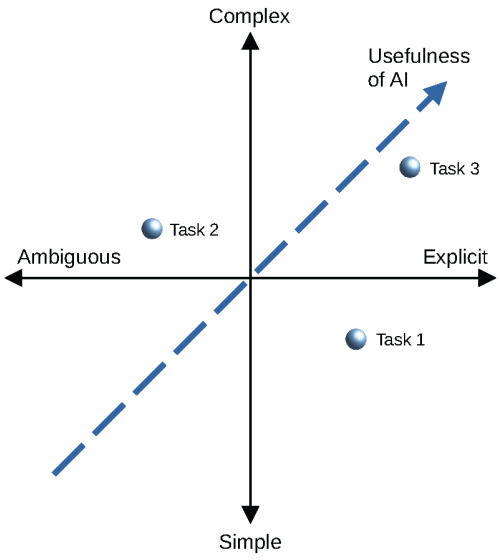
While we acknowledge that Turnitin's AI detection tool cannot definitively prove the extent to which AI was used by respondents — and the scores are almost certainly an underestimate, unless most people in the treatment group reported using AI without actually doing so — these results give us confidence that both groups largely complied with instructions on AI use.

E Tasks and Grading Scheme

After answering the self-assessment questions, the participants were informed that they would now start their first task. Depending on the group they were in, they were told to complete the tasks with/without using ChatGPT. Participants were also reminded that they may earn a substantial bonus if they perform well and that we expect them to take approximately four minutes for each task. Figure S8 shows how the tasks map onto a two-dimensional space distinguishing between complexity and ambiguity. Figure S9 to Figure S18 show screenshots of the instructions participants received and the tasks.

The grading scheme works as follows. ChatGPT (GPT-4) is prompted with the evaluation prompt and the relevant text for the task. The next prompt consists of a number of responses, identified with the respondent ID and clearly separated from one another. ChatGPT then provides a score and a short justification for the score for each answer. The scores are recorded in an Excel file. This process is repeated iteratively. Periodically, ChatGPT has to be re-prompted with the initial prompt and text, owing to its limited context window.

Figure S8: Usefulness of AI in relation to task complexity and ambiguity



Note: Schematic depiction of the usefulness of AI in relation to task complexity and ambiguity. The potential for AI to improve performance increases as we move from simple and ambiguous tasks to complex and explicit ones. Approximate placement of our tasks in the coordinate system for illustrative purposes.

Figure S9: Instructions (Control Group)

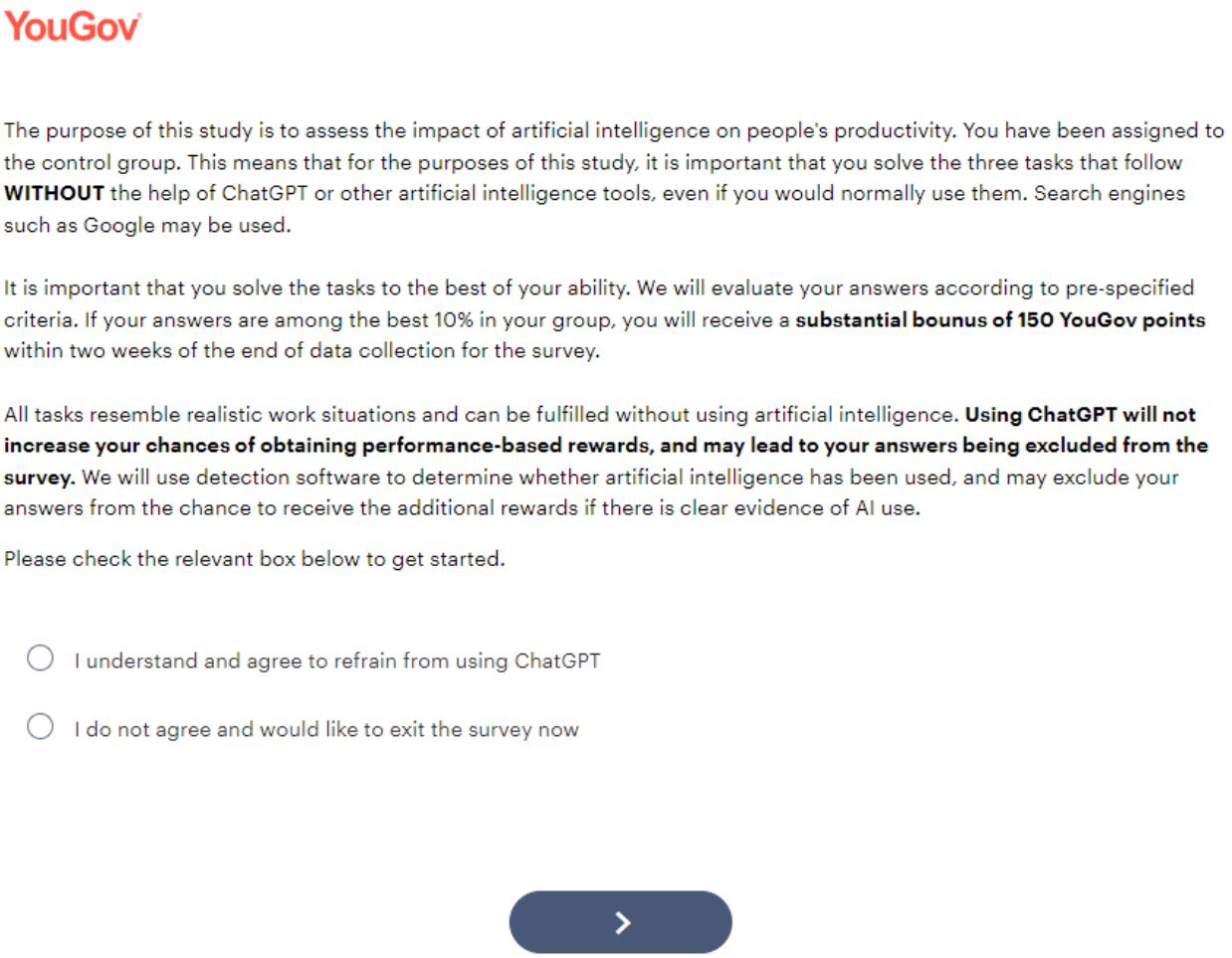


Figure S10: Instructions (Treatment Group)

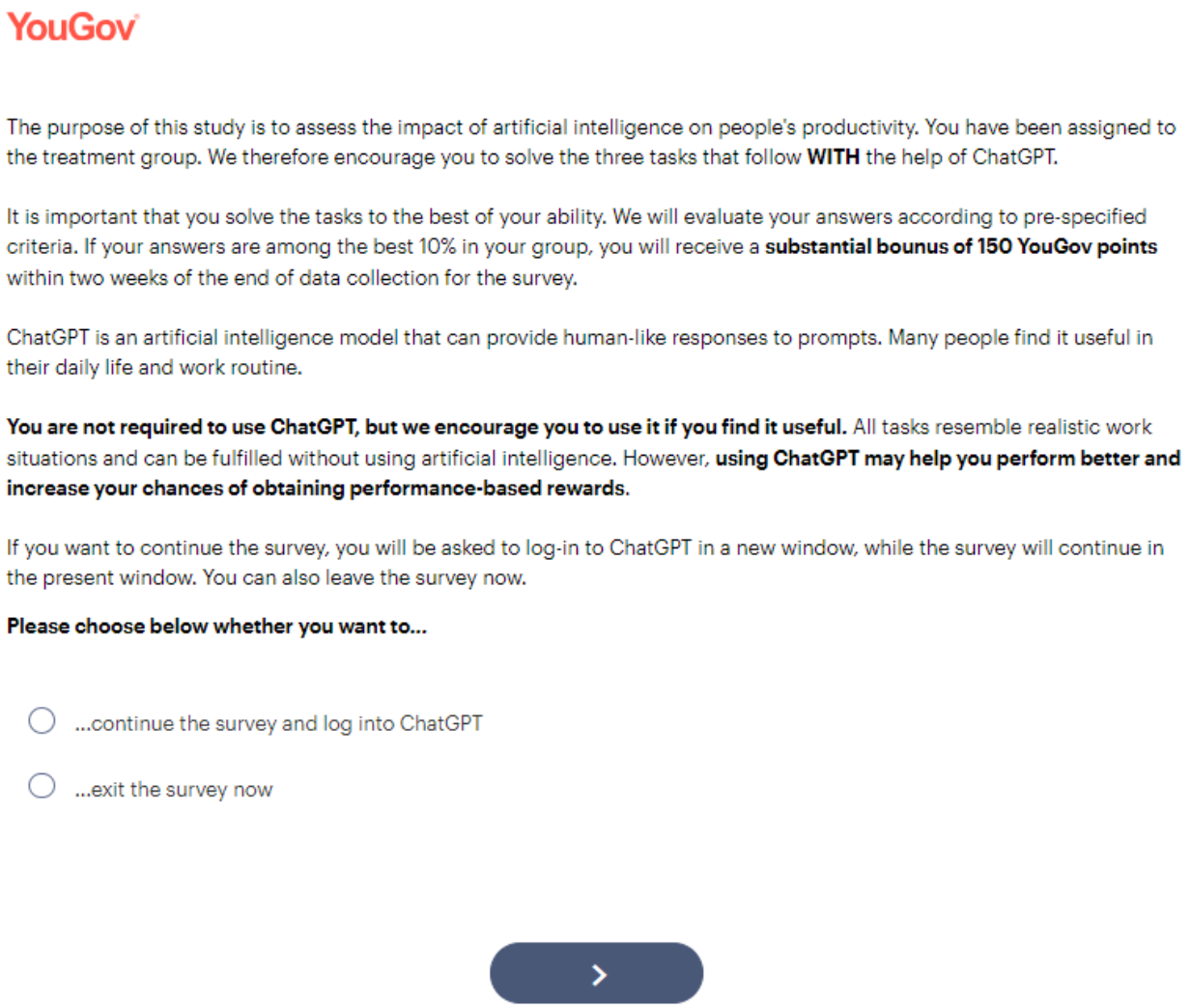


Figure S11: ChatGPT login (Treatment Group)

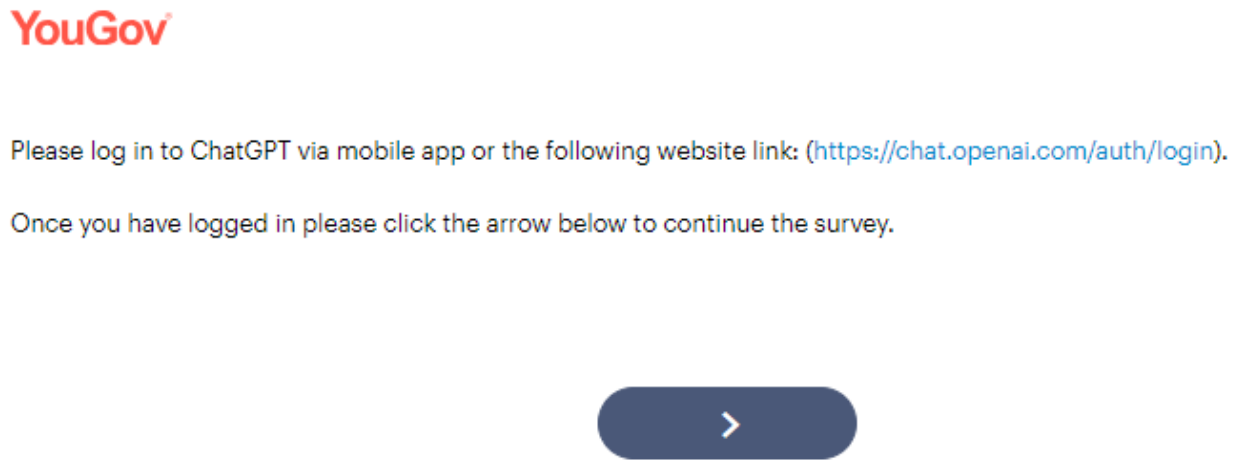


Figure S12: Instructions reminder (Control Group)

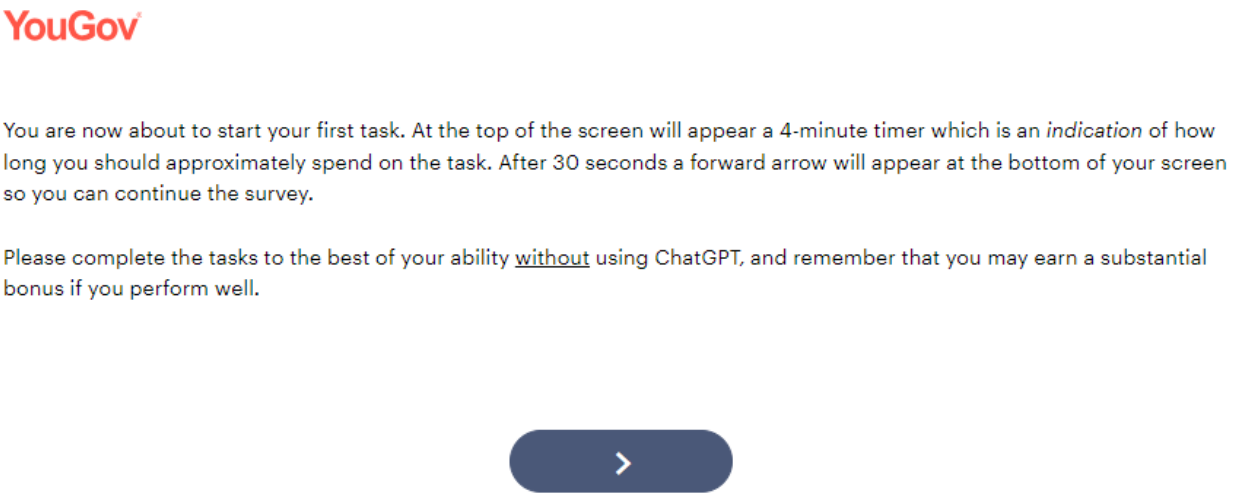


Figure S13: Instructions reminder (Treatment Group)

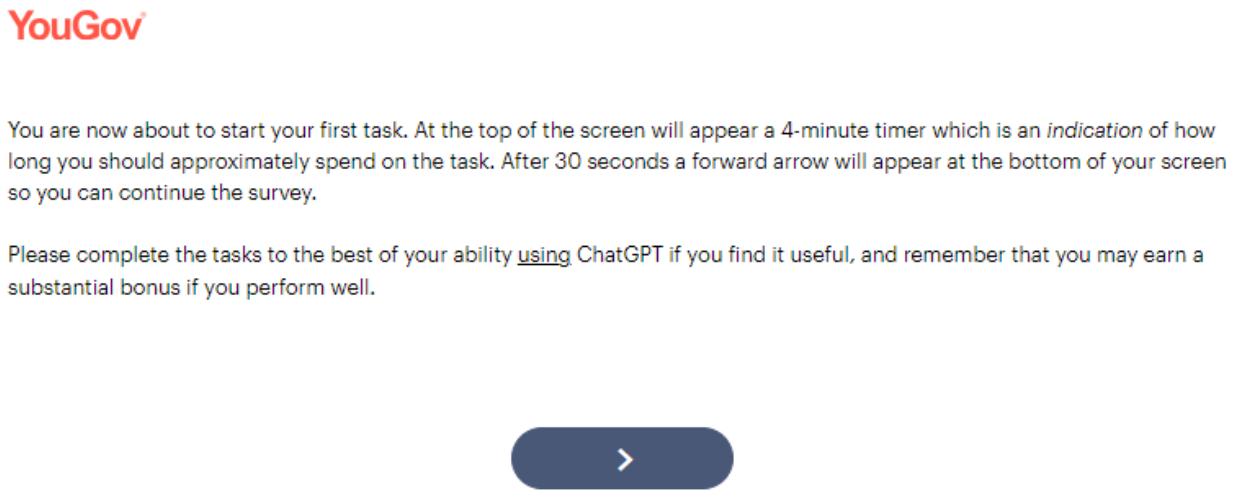


Figure S14: Email task

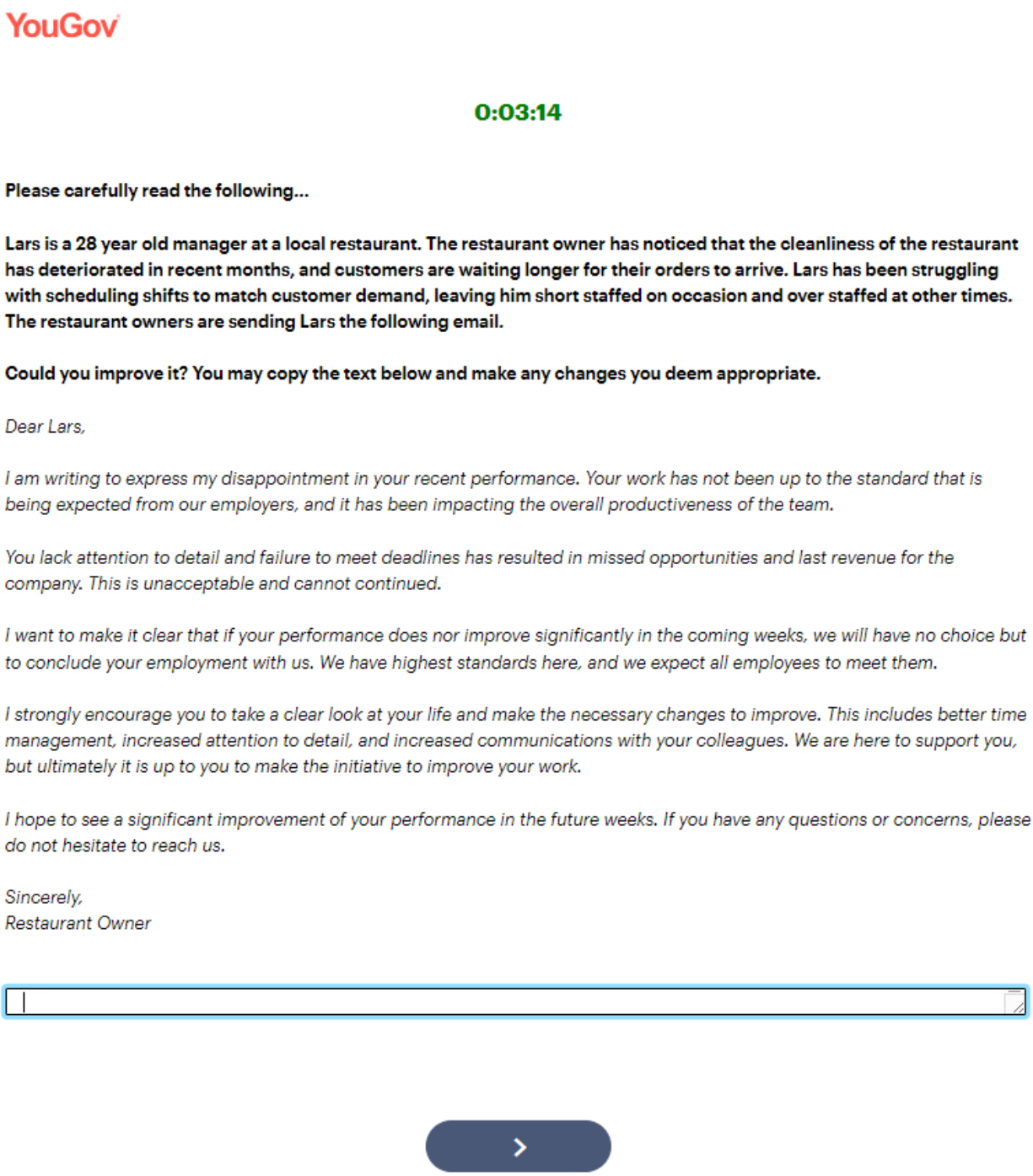


Figure S15: Assessment task (part 1)

YouGov

0:03:54

Please carefully read the following... The following two texts, adapted from a UK newspaper, take opposing views on the question of whether a universal basic income (UBI) is a good idea.

Excerpt 1:
"The idea of a UBI has made recurrent appearances in history. This time, though, it is likely to have greater staying power, as the prospect of sufficient income from jobs grows bleaker for the poor and less educated.

UBI is a somewhat uneasy mix of two objectives: poverty relief and the rejection of work as the defining purpose of life. The first is political and practical; the second is philosophical or ethical.

The main argument for UBI as poverty relief is the inability of available paid work to guarantee a secure and decent existence for all. In the industrial age, factory work became the only source of income for most people – a source that was interrupted by periodical bouts of unemployment. The labour movement responded by demanding "work or maintenance". This led to the creation of a system of social security, "welfare capitalism", designed to provide people an income during enforced interruptions of work. Soon, the idea of interruption from work was extended to include the disabled and women bringing up children.

In the 1980s, conservative governments unwittingly extended the scope of welfare further, as they dismantled institutions and legislation designed to protect wages and jobs. "In work" benefits, were introduced to enable employed workers to earn a "living wage". At the same time, governments started to cut back on welfare entitlements. In this newly precarious environment of work and welfare, UBI is seen as guaranteeing the basic income previously promised by work and welfare, but no longer reliably secured by either.

The ethical case for UBI is different. Its source is the idea, found both in the Bible and in classical economics, that work is a curse or "cost", undertaken only for the sake of making a living. As technological innovation causes per capita income to rise, people will need to work less to satisfy their needs. UBI provides a practical path to navigate this transition. Most of the hostility to UBI has come when it is stated in this second form. But to argue that an income independent of the job market is bound to be demoralising is historically inaccurate. If it were true, we would want to abolish all inherited income.

A standard objection to UBI is that it is unaffordable. This partly depends on what parameters are set: the UBI's level; which benefits (if any) it replaces; whether only citizens or all residents are eligible; and so on. But this is not the main point. The overwhelming evidence is that the lion's share of productivity gains in the last 30 years has gone to the very wealthy. Even a partial reversal of this long regressive trend for wealth and income would fund a modest initial basic income

Unless we change our system of income generation, there will be no way to check the concentration of wealth in the hands of the rich. A UBI that grows in line with capital productivity would ensure that the benefits of growth go to the many, not just to the few."

Figure S16: Assessment task (part 2)

Excerpt 2:

"A recent study sheds doubt on ambitious claims made for a UBI, the scheme that would give everyone regular, unconditional cash payments that are enough to live on. Its advocates claim it would help to reduce poverty, and narrow inequalities.

New research reviewed for the first time 16 projects that have tested different ways of distributing regular cash payments to individuals across a range of countries, as well as copious literature on the topic. It could find no evidence to suggest that such a scheme could be sustained for all individuals in any country – or that this approach could achieve lasting improvements in wellbeing or equality. The research confirms the importance of generous income support, but everything turns on how much money is paid, under what conditions and with what consequences for the welfare system.

Cash payment schemes around the world have been claimed to show that UBI "works". In fact, what's been tested in practice is almost infinitely varied. The Alaska Permanent Fund pays all adults and children a dividend each year – in 2018, it was $1,600 (£1,230). The scheme is popular and enduring; but it makes no claim to sufficiency and has done nothing to reduce child poverty or to prevent widening income inequalities. Finland undertook a two-year trial of modest monthly payments of €560 (£477) to 2,000 unemployed people – but the government has refused to fund further expansion. It told us little about UBI except that, when push comes to shove, elected politicians may balk at paying for a universal scheme.

The cost of a sufficient UBI scheme would be extremely high according to the International Labour Office, which estimates average costs equivalent to 20-30% of GDP in most countries. Costs can be reduced by paying smaller amounts to fewer individuals. But there is no evidence that a partial or conditional UBI could do anything to mitigate, let alone reverse, current trends towards worsening poverty and inequality.

Costs may be offset by raising taxes or shifting expenditure, but either way there are huge trade-offs. Money spent on cash payments cannot be invested elsewhere. As the report observes, "If cash payments are allowed to take precedence, there's a serious risk of [...] setting a pattern for future development that promotes commodification rather than emancipation." This may help to explain why UBI has attracted support from Silicon Valley tycoons, who are more interested in defending consumer capitalism than in tackling poverty and inequality.

The report concludes that the money needed to pay for an adequate UBI scheme "would be better spent on reforming social protection systems, and building more and better-quality public services". Collective provision offers more cost-effective, socially just, redistributive and sustainable ways of meeting people's needs."

Please provide a short assessment (no more than 1200 characters) of which of the texts is more convincing. Do not simply state which text you agree with more, but provide a reasoned assessment as to why one of the texts is more convincing than the other.

There is no right or wrong answer, people may have valid reasons for finding either text more convincing. You may answer in bullet points.

Figure S17: Comprehension task (part 1)

YouGov

0:03:53

Please carefully read the following...

"In tracing the changing face of the Irish landscape, scholars have traditionally relied primarily on evidence from historical documents. However, such documentary sources provide a fragmentary record at best. Reliable accounts are very scarce for many parts of Ireland prior to the seventeenth century, and many of the relevant documents from the sixteenth and seventeenth centuries focus selectively on matters relating to military or commercial interests.

Studies of fossilized pollen grains preserved in peats and lake muds provide an additional means of investigating vegetative landscape change. Details of changes in vegetation resulting from both human activities and natural events are reflected in the kinds and quantities of minute pollen grains that become trapped in sediments. Analysis of samples can identify which kinds of plants produced the preserved pollen grains and when they were deposited, and in many cases the findings can serve to supplement or correct the documentary record.

For example, analyses of samples from Long Lough in County Down have revealed significant patterns of cereal-grain pollen beginning by about 400 A.D. The substantial clay content of the soil in this part of Down makes cultivation by primitive tools difficult. Historians thought that such soils were not tilled to any significant extent until the introduction of the moldboard plough to Ireland in the seventh century A.D. Because cereal cultivation would have required tilling of the soil, the pollen evidence indicates that these soils must indeed have been successfully tilled before the introduction of the new plough.

Another example concerns flax cultivation in County Down, one of the great linen-producing areas of Ireland during the eighteenth century. Some aspects of linen production in Down are well documented, but the documentary record tells little about the cultivation of flax, the plant from which linen is made, in that area. The record of eighteenth-century linen production in Down, together with the knowledge that flax cultivation had been established in Ireland centuries before that time, led some historians to surmise that this plant was being cultivated in Down before the eighteenth century. But pollen analyses indicate that this is not the case; flax pollen was found only in deposits laid down since the eighteenth century.

It must be stressed, though, that there are limits to the ability of the pollen record to reflect the vegetative history of the landscape. For example, pollen analyses cannot identify the species, but only the genus or family, of some plants. Among these is madder, a cultivated dye plant of historical importance in Ireland. Madder belongs to a plant family that also comprises various native weeds, including goosegrass. If madder pollen were present in a deposit it would be indistinguishable from that of uncultivated native species."

Figure S18: Comprehension task (part 2)

Please complete the sentences below so that they accurately reflect the information contained in the text.

Regarding the cultivation of cereal grains, the passage indicates that pollen analyses have provided evidence against...

The author's likely position on the limits of historical records is...

In the structure of the author's argument, the relationship between the second paragraph and the final paragraph of the passage is...

Email Task

ChatGPT evaluation prompt: Below is a draft of an email. I want you to remember it. In the following, I will show you revised versions of the email. Evaluate whether the revised versions constitute an improvement, considering the following criteria: language (Have all or almost all spelling and grammar mistakes been corrected? Is the language clear and concise, without adding redundant information?) and tone (The revised version should be more constructive and not directly threaten termination).
Apply a 5-point scale, where the levels mean the following:

5/5: Noticeable improvement on both dimensions

4/5: Noticeable improvement only on one dimension (e.g., corrected all or almost all spelling and grammar mistakes but no improvement in the tone of the email)

3/5: No improvement on either dimension (this includes cases where only a minority of the mistakes have been corrected) OR improvement in one dimension counterbalanced by worsening in the other (e.g., corrected all or almost all spelling and grammar mistakes but adopted even harsher tone)

2/5: Revised version is worse than the draft on one dimension and unchanged on the other (e.g., spelling mistakes have not been corrected and tone is even harsher)

1/5: Revised version is worse than the draft on both dimensions, or is a non-answer such as N/A.

There are no half-grades (e.g., 4.5/5). Ignore encoding errors. "|||||||" denotes a line break.

Original draft:

[Email text here]

The revised versions to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID $\sim$ x/5 $\sim$ comments.

Assessment Task

ChatGPT evaluation prompt: Below are two texts on the merits of universal basic income (UBI). Remember them. In the next prompts, they will be followed by assessments which text is more convincing. Evaluate the assessments. There is no right or wrong answer, people may have valid reasons for finding either text more convincing. The evaluation should focus on how well-reasoned the assessment is.
Apply a 5-point scale, where the levels mean the following:

5/5: The answer provides a comprehensive and well-reasoned appraisal of strengths and weaknesses of both texts. It includes a plausible judgment which text is more convincing which follows from the appraisal.

4/5: The answer provides a well-reasoned appraisal of strengths and weaknesses of both texts, but it lacks detail. The judgment which text is more convincing follows from the incomplete appraisal of the texts.

3/5: The appraisal of strengths and weaknesses of both texts is not well-reasoned or comprehensive (for example, it only discusses the strengths of one text) and/or the judgment which text is more convincing does not follow from the appraisal.

2/5: The answer is an unsupported statement which does not discuss the strengths and weaknesses of the texts, or simply summarises the texts.

1/5: The answer is unrelated to the texts, or is unintelligible, or comments on the merits of UBI itself rather than on the quality of the texts, or is a non-answer such as N/A.

There are no half-grades (e.g., 4.5/5). Ignore encoding errors. "|||||||" denotes a line break.

[Text 1 here]

[Text 2 here]

The assessments to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID $\sim$ x/5 $\sim$ comments.

Text Comprehension Questions

Pollen Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements. Each of the statements is meant to complete the sentence "Regarding the cultivation of cereal grains, the passage indicates that pollen analyses have provided evidence against…". The key takeaway from the passage is that pollen analyses have provided evidence against the view that in certain parts of County Down, cereal grains were not cultivated to any significant extent before the seventh century, when the moldboard plough was introduced.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

5/5: The answer correctly identifies the key takeaway and mentions cereal grains.

3/5: The answer correctly identifies one of the tangential points, such as that pollen evidence challenges the view that soil was not tilled in Country Down before the seventh century or that pollen evidence may challenge documentary evidence.

1/5: The answer misrepresents the information in the text, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comment (max. 1 sentence). Separate the evaluations by an empty line.

Historical Records Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements. Each of the statements is meant to complete the sentence "The author's likely position on the limits of historical records is…". The key takeaway is that the author believes that while historical documents are valuable for tracing the past, their record is often fragmentary and focuses selectively on certain aspects. Thus, the author believes these documents should be supplemented with other investigative techniques to provide a more comprehensive understanding of history.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

5/5: The answer correctly identifies a) the author's likely view of historical records as fragmentary and/or selective and b) the author's likely view that additional investigative techniques should be used to supplement or correct the documentary record.

3/5: The answer correctly identifies either a) the author's likely view of historical records as fragmentary and/or selective or b) the author's likely view that additional investigative techniques should be used to supplement or correct the documentary record.

1/5: The answer misrepresents the author's likely position, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID ∼ x/5 ∼ comment (max. 1 sentence). Separate the evaluations by an empty line.

Paragraph Structure Question

ChatGPT evaluation prompt: Below is a text passage, remember it. It will be followed by different statements about the relationship between the second and the last paragraph of the text. Each of the statements is meant to complete the sentence "In the structure of the author's argument, the relationship between the second paragraph and the final paragraph of the passage is...". The important takeaway is that the final paragraph qualifies the claim made in the second paragraph.

Evaluate the accuracy of the statements on a scale from 1 to 5 according to the following grading scheme:

> 5/5: The answer correctly identifies that the final paragraph qualifies the claim made in the second paragraph.
>
> 3/5: The answer correctly identifies that the author discusses the limitations of pollen analysis, but fails to make explicit the argumentative structure of the text and relationship between the paragraphs.
>
> 1/5: The answer misrepresents the relationship between the paragraphs, or is unrelated to the question, or is unintelligible, or is a non-answer such as N/A.

Do not use scores other than 1, 3, and 5. Ignore encoding errors.

[Text here]

The statements to evaluate will be below, in the format "answer ID": "answer text". Provide output in the following format: answer ID $\sim$ x/5 $\sim$ comment (max. 1 sentence). Separate the evaluations by an empty line.

F Supplementary Analyses

Table S18: Within-occupational group variation in performance

Occ. type	Task	CV Treated	CV Control	F-statistic	p-value	Emp. share
Professionals	Email	34.072	36.407	1.142	0.171	39.7
	Assess.	**40.196**	33.423	1.446	0.004	
	Comp.	44.391	48.751	1.206	0.090	
Managers	Email	37.391	37.462	1.003	0.492	27.7
	Assess.	40.317	37.156	1.177	0.164	
	Comp.	46.790	51.613	1.217	0.123	
Clerical & Service	Email	31.771	**39.273**	1.528	0.013	22.6
	Assess.	40.574	36.275	1.251	0.113	
	Comp.	44.325	47.075	1.128	0.263	
Manual	Email	60.384	46.716	1.671	0.100	6.4
	Assess.	**56.921**	39.038	2.126	0.030	
	Comp.	**69.474**	44.183	2.473	0.012	
Other	Email	25.590	**46.810**	3.346	0.009	3.7
	Assess.	31.180	32.856	1.110	0.421	
	Comp.	44.525	52.980	1.416	0.242	

Note: Coefficient of variation (CV) of task quality, by occupational group. Numbers in bold signify significantly higher CV than in the other group. AI tends to reduce variation in performance in all groups except manual workers in the email and comprehension tasks, and increases variation for all groups except "Other" in the assessment task.

Table S19: ITT estimates

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Unconditional treatment effects						
Treatment	−89.212***	−156.995***	−63.840***	0.840***	0.431***	1.348***
	(9.790)	(13.084)	(11.439)	(0.081)	(0.064)	(0.081)
Constant	304.234***	355.858***	317.427***	3.078***	2.428***	2.028***
	(6.812)	(9.104)	(7.959)	(0.056)	(0.045)	(0.057)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.074	0.122	0.029	0.094	0.042	0.209
Panel B: Interaction treatment x compliance						
Treatment	−41.831**	−44.803*	−36.078*	−0.063	−0.129	−0.040
	(15.286)	(19.457)	(16.611)	(0.123)	(0.096)	(0.104)
AI Use	−102.666***	−148.783***	−143.619***	−0.903***	−0.560**	0.088
	(25.210)	(38.186)	(34.256)	(0.202)	(0.189)	(0.214)
Treatment x AI Use	28.252	−17.611	91.493**	2.027***	1.289***	1.940***
	(29.750)	(43.225)	(38.470)	(0.239)	(0.214)	(0.240)
Constant	312.072***	364.170***	325.450***	3.147***	2.460***	2.023***
	(6.966)	(9.026)	(8.097)	(0.056)	(0.045)	(0.050)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.107	0.187	0.053	0.173	0.096	0.407
Panel C: Unconditional treatment effects excluding non-compliers						
Treatment	−116.245***	−211.197***	−88.204***	1.061***	0.601***	1.989***
	(10.466)	(13.946)	(12.315)	(0.084)	(0.069)	(0.080)
Constant	312.072***	364.170***	325.450***	3.147***	2.460***	2.023***
	(6.862)	(9.023)	(7.843)	(0.055)	(0.045)	(0.051)
Observations	870	872	853	870	872	853
R^2	0.124	0.209	0.057	0.155	0.080	0.418

Note: Figure S7 is created based on the models in Panels B. Panel C shows that excluding non-compliers increases effect sizes. *p<0.05; **p<0.01; ***p<0.001.

Table S20: ITT estimates for heterogeneous effects analyses

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	-85.635^{***}	-144.678^{***}	-55.944^{***}	0.819^{***}	0.288^{***}	1.174^{***}
	(11.717)	(15.662)	(13.720)	(0.097)	(0.077)	(0.097)
GPT frequent	-32.496^{**}	-25.621	-15.089	-0.144	-0.148	-0.047
	(15.980)	(21.360)	(18.712)	(0.132)	(0.104)	(0.132)
Treatment x GPT frequent	1.307	-25.266	-16.657	0.108	0.446^{***}	0.497^{***}
	(21.640)	(28.927)	(25.340)	(0.179)	(0.141)	(0.179)
Constant	311.919^{***}	361.917^{***}	320.996^{***}	3.112^{***}	2.463^{***}	2.039^{***}
	(7.771)	(10.388)	(9.100)	(0.064)	(0.051)	(0.064)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.082	0.129	0.033	0.096	0.052	0.220
Treatment	-119.137^{***}	-179.803^{***}	-63.848^{*}	0.540^{**}	0.347^{*}	0.947^{***}
	(21.739)	(28.885)	(25.396)	(0.180)	(0.143)	(0.180)
Under 50	-59.251^{***}	-93.997^{***}	-51.780^{*}	-0.064	0.096	0.067
	(17.403)	(23.124)	(20.331)	(0.144)	(0.114)	(0.144)
Treatment × Under 50	35.910	25.498	-1.975	0.381	0.110	0.514^{*}
	(24.319)	(32.312)	(28.410)	(0.201)	(0.160)	(0.202)
Constant	352.451^{***}	432.351^{***}	359.565^{***}	3.130^{***}	2.350^{***}	1.973^{***}
	(15.700)	(20.860)	(18.340)	(0.130)	(0.103)	(0.130)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.086	0.143	0.042	0.099	0.046	0.222
Treatment	-93.433^{***}	-166.938^{***}	-77.178^{***}	0.926^{***}	0.448^{***}	1.325^{***}
	(11.895)	(15.898)	(13.888)	(0.098)	(0.078)	(0.099)
Female	4.576	-13.900	-17.172	0.247^{*}	0.002	0.039
	(14.649)	(19.580)	(17.104)	(0.121)	(0.096)	(0.122)
Treatment x Female	12.741	30.833	41.309	-0.270	-0.053	0.069
	(20.959)	(28.013)	(24.471)	(0.173)	(0.137)	(0.174)
Constant	302.785^{***}	360.258^{***}	322.863^{***}	3.000^{***}	2.428^{***}	2.015^{***}
	(8.242)	(11.017)	(9.624)	(0.068)	(0.054)	(0.068)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.075	0.123	0.032	0.098	0.042	0.210
Treatment	-82.562^{***}	-131.173^{***}	-53.678^{*}	0.728^{***}	0.428^{**}	1.321^{***}
	(21.767)	(28.944)	(25.337)	(0.178)	(0.141)	(0.179)
Uni	4.230	3.598	6.371	0.181	0.160	0.264
	(17.066)	(22.692)	(19.864)	(0.140)	(0.111)	(0.140)
Treatment x Uni	-7.644	-33.169	-13.461	0.146	0.012	0.030
	(24.451)	(32.513)	(28.461)	(0.200)	(0.159)	(0.201)
Constant	301.123^{***}	353.965^{***}	313.569^{***}	2.936^{***}	2.303^{***}	1.820^{***}
	(15.210)	(20.225)	(17.704)	(0.125)	(0.099)	(0.125)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.073	0.124	0.030	0.101	0.047	0.214
Treatment	-84.341^{***}	-132.305^{***}	-71.796^{***}	0.851^{***}	0.389^{***}	1.420^{***}
	(17.176)	(22.941)	(20.071)	(0.141)	(0.112)	(0.142)
Professional	11.502	10.584	-10.424	0.152	0.174	0.275^{*}
	(14.401)	(19.234)	(16.828)	(0.119)	(0.094)	(0.119)
Treatment x Professional	-7.522	-36.356	11.975	-0.021	0.054	-0.115
	(20.914)	(27.934)	(24.440)	(0.172)	(0.136)	(0.173)
Constant	296.630^{***}	348.861^{***}	324.318^{***}	2.978^{***}	2.313^{***}	1.846^{***}
	(11.709)	(15.639)	(13.683)	(0.096)	(0.076)	(0.097)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.075	0.123	0.029	0.097	0.050	0.214

Note: *p<0.05; **p<0.01; ***p<0.001.

Table S21: ITT analyses with weights

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−86.298***	−156.017***	−65.938***	0.810***	0.446***	1.424***
	(9.925)	(13.113)	(11.491)	(0.082)	(0.064)	(0.082)
Constant	296.744***	347.756***	312.276***	3.068***	2.414***	1.988***
	(6.873)	(9.081)	(7.958)	(0.056)	(0.044)	(0.057)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.068	0.120	0.031	0.087	0.045	0.226
Treatment	−129.860***	−187.072***	−73.030*	0.494*	0.366*	0.958***
	(26.407)	(34.775)	(30.604)	(0.218)	(0.171)	(0.218)
Under 50	−68.258***	−97.049***	−55.518*	−0.017	0.114	0.042
	(20.470)	(26.958)	(23.724)	(0.169)	(0.133)	(0.169)
Treatment × Under 50	49.422	33.810	6.741	0.372	0.099	0.552*
	(28.476)	(37.501)	(33.003)	(0.235)	(0.184)	(0.235)
Constant	356.261***	432.377***	360.684***	3.082***	2.315***	1.951***
	(19.115)	(25.172)	(22.153)	(0.158)	(0.124)	(0.158)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.079	0.136	0.040	0.091	0.048	0.235
Treatment	−88.236***	−159.058***	−72.931***	0.863***	0.459***	1.406***
	(11.621)	(15.365)	(13.460)	(0.096)	(0.075)	(0.096)
Female	12.215	2.567	−5.273	0.164	0.057	0.130
	(15.595)	(20.620)	(18.063)	(0.128)	(0.101)	(0.129)
Treatment x Female	6.441	10.875	25.519	−0.198	−0.050	0.059
	(22.347)	(29.546)	(25.883)	(0.184)	(0.144)	(0.184)
Constant	293.520***	347.079***	313.668***	3.024***	2.399***	1.954***
	(8.011)	(10.592)	(9.279)	(0.066)	(0.052)	(0.066)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.070	0.120	0.032	0.088	0.045	0.228
Treatment	−66.839***	−106.135***	−37.459	0.683***	0.518***	1.503***
	(19.235)	(25.259)	(22.151)	(0.157)	(0.123)	(0.156)
Uni	15.847	25.856	27.676	0.152	0.179	0.412***
	(15.243)	(20.017)	(17.554)	(0.124)	(0.097)	(0.124)
Treatment x Uni	−25.873	−68.680**	−39.357	0.174	−0.090	−0.117
	(22.569)	(29.637)	(25.990)	(0.184)	(0.144)	(0.183)
Constant	285.749***	330.370***	293.855***	2.961***	2.289***	1.698***
	(12.795)	(16.803)	(14.735)	(0.104)	(0.082)	(0.104)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.067	0.125	0.033	0.095	0.051	0.240
Treatment	−75.754***	−116.239***	−59.637**	0.803***	0.408***	1.571***
	(16.303)	(21.499)	(18.889)	(0.134)	(0.105)	(0.134)
Professional	17.600	31.638	5.618	0.080	0.121	0.357**
	(14.048)	(18.525)	(16.276)	(0.116)	(0.090)	(0.115)
Treatment x Professional	−17.507	−63.466*	−10.098	0.004	0.050	−0.251
	(20.567)	(27.122)	(23.829)	(0.169)	(0.132)	(0.169)
Constant	286.139***	328.694***	308.891***	3.019***	2.341***	1.773***
	(10.904)	(14.379)	(12.634)	(0.090)	(0.070)	(0.090)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.069	0.125	0.031	0.087	0.049	0.233

Note: *p<0.05; **p<0.01; ***p<0.001.

Table S22: ITT analysis with controls

	Dependent variable:					
	Email Time	Ass. Time	Comp.Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	−89.355***	−157.988***	−64.704***	0.840***	0.436***	1.344***
	(9.892)	(13.135)	(11.510)	(0.081)	(0.064)	(0.081)
Under 40	−26.170**	−46.369***	−35.266**	0.059	0.220***	0.339***
	(10.046)	(13.339)	(11.688)	(0.082)	(0.065)	(0.082)
Female	14.727	5.795	6.450	0.102	−0.040	0.026
	(10.694)	(14.201)	(12.443)	(0.088)	(0.069)	(0.087)
Degree	−1.079	−7.533	4.211	0.207*	0.094	0.191
	(12.749)	(16.929)	(14.834)	(0.105)	(0.082)	(0.104)
Professional	5.207	−10.983	−10.539	0.110	0.204**	0.222*
	(11.093)	(14.730)	(12.907)	(0.091)	(0.072)	(0.091)
Constant	311.031***	392.709***	338.819***	2.778***	2.117***	1.543***
	(14.058)	(18.667)	(16.356)	(0.115)	(0.091)	(0.115)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.080	0.133	0.038	0.103	0.063	0.230

Note: *p<0.05; **p<0.01; ***p<0.001.

Table S23: LATE analysis with weights

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−128.419***	−223.851***	−99.492***	1.205***	0.640***	2.149***
	(14.497)	(18.284)	(17.228)	(0.123)	(0.092)	(0.113)
Constant	307.323***	361.309***	318.341***	2.968***	2.376***	1.857***
	(7.623)	(9.625)	(8.667)	(0.065)	(0.048)	(0.057)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.102	0.169	0.043	0.062	0.050	0.351
Panel B: First stage						
Treatment	0.672***	0.697***	0.663***	0.672***	0.697***	0.663***
	(0.022)	(0.021)	(0.022)	(0.022)	(0.021)	(0.022)
Constant	0.082***	0.061***	0.061***	0.082***	0.061***	0.061***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.468	0.507	0.466	0.468	0.507	0.466
F Statistic (df = 1; 1039)	913.439***	1,070.277***	906.232***	913.439***	1,070.277***	906.232***

Note: Replicating Table S1 using survey weights. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S24: LATE analysis with controls

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−133.717***	−236.270***	−103.046***	1.257***	0.652***	2.141***
	(14.679)	(19.151)	(18.236)	(0.122)	(0.095)	(0.117)
Female	9.756	−5.501	1.739	0.149	−0.009	0.124
	(10.611)	(13.861)	(12.395)	(0.088)	(0.069)	(0.080)
Under 40	−12.015	−22.672	−21.409	−0.074	0.154*	0.052
	(10.059)	(13.117)	(11.849)	(0.084)	(0.065)	(0.076)
Uni degree	3.621	−2.488	9.539	0.163	0.080	0.080
	(12.646)	(16.504)	(14.776)	(0.105)	(0.082)	(0.095)
Professional	9.578	−0.908	−8.708	0.069	0.176*	0.184*
	(11.026)	(14.409)	(12.855)	(0.092)	(0.072)	(0.083)
Constant	308.357***	386.059***	333.392***	2.803***	2.136***	1.656***
	(13.844)	(18.024)	(15.976)	(0.115)	(0.089)	(0.103)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.095	0.176	0.048	0.092	0.074	0.362
Panel B: First stage						
Treatment	0.668***	0.669***	0.628***	0.668***	0.669***	0.628***
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
Female	−0.037	−0.048*	−0.046	−0.037	−0.048*	−0.046
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Under 40	0.106***	0.100***	0.134***	0.106***	0.100***	0.134***
	(0.022)	(0.022)	(0.023)	(0.022)	(0.022)	(0.023)
Uni degree	0.035	0.021	0.052	0.035	0.021	0.052
	(0.029)	(0.028)	(0.029)	(0.029)	(0.028)	(0.029)
Professional	0.033	0.043	0.018	0.033	0.043	0.018
	(0.025)	(0.024)	(0.025)	(0.025)	(0.024)	(0.025)
Constant	−0.020	−0.028	−0.053	−0.020	−0.028	−0.053
	(0.031)	(0.031)	(0.032)	(0.031)	(0.031)	(0.032)
Observations	1,027	1,027	1,027	1,027	1,027	1,027
R^2	0.478	0.487	0.447	0.478	0.487	0.447
F Statistic (df = 5; 1021)	186.970***	193.739***	164.906***	186.970***	193.739***	164.906***

Note: Analysis in Table S1 with added controls. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.

Table S25: LATE interaction with age - under 50/50+

	Dependent variable:					
	Email Time	Ass. Time	Comp. Time	Email Score	Ass. Score	Comp. Score
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Second stage						
Used AI	−233.015***	−307.090***	−131.533*	1.056**	0.593*	1.951***
	(42.518)	(48.326)	(52.137)	(0.356)	(0.242)	(0.338)
Under 50	−56.056**	−85.870***	−46.604*	−0.145	0.066	−0.057
	(18.681)	(23.705)	(21.004)	(0.156)	(0.119)	(0.136)
Used AI × Under 50	115.879*	84.392	33.710	0.241	0.067	0.220
	(45.201)	(52.484)	(55.443)	(0.379)	(0.263)	
Constant	359.441***	438.492***	360.880***	3.098***	2.338***	1.954***
	(16.644)	(21.142)	(18.657)	(0.139)	(0.106)	(0.121)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.086	0.178	0.049	0.078	0.060	0.354
Panel B: First stage						
Treatment	0.511***	0.586***	0.485***	0.511***	0.586***	0.485***
	(0.049)	(0.048)	(0.049)	(0.049)	(0.048)	(0.049)
Under 50	0.057	0.044	0.056	0.057	0.044	0.056
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)
Treatment × Under 50	0.199***	0.107*	0.187***	0.199***	0.107*	0.187***
	(0.055)	(0.054)	(0.055)	(0.055)	(0.054)	(0.055)
Constant	0.030	0.020	0.010	0.030	0.020	0.010
	(0.035)	(0.035)	(0.036)	(0.035)	(0.035)	(0.036)
Observations	1,041	1,041	1,041	1,041	1,041	1,041
R^2	0.485	0.482	0.453	0.485	0.482	0.453
F Statistic (df = 3; 1037)	325.819***	322.176***	285.929***	325.819***	322.176***	285.929***

Note: Replication of Table S4 with the cutoff at 50 years instead of 40 years. The second stage is the effect of reported AI use on productivity. The first stage is the effect of treatment assignment on reported AI use. *p<0.05; **p<0.01; ***p<0.001.