

Mourelatos, Evangelos; Zervas, Panagiotis; Lagios, Dimitris; Tzimas, Giannis

Working Paper

Can AI Bridge the Gender Gap in Competitiveness?

GLO Discussion Paper, No. 1404

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Mourelatos, Evangelos; Zervas, Panagiotis; Lagios, Dimitris; Tzimas, Giannis (2024) : Can AI Bridge the Gender Gap in Competitiveness?, GLO Discussion Paper, No. 1404, Global Labor Organization (GLO), Essen

This Version is available at:
<http://hdl.handle.net/10419/284850>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

Can AI Bridge the Gender Gap in Competitiveness?

Evangelos Mourelatos^{1,2}, Panagiotis Zervas³, Dimitris Lagios³ and Giannis Tzimas³

¹ *Department of Economics, Accounting and Finance, Oulu Business School, University of Oulu, Oulu, Finland*

² *Global Labour Organization, Essen, Germany*

³ *Data and Media Laboratory, Department of Electrical and Computer Engineering, University of Peloponnese, Patras, Greece*

Abstract

This paper employs an online real-effort experiment to investigate gender disparities in the selection of individuals into competitive working environments when assisted by artificial intelligence (AI). In contrast to previous research suggesting greater competitiveness among men, our findings reveal that both genders are equally likely to compete in the presence of AI assistance. Surprisingly, the introduction of AI eliminates an 11-percentage-point gender gap, between men and women in our competitive scenario. We also discuss how the gender gap in tournament entry appears to be contingent on ChatGPT selection rather than being omnipresent. Notably, 47% of female participants independently chose to utilize ChatGPT, while 55% of males did the same. However, when ChatGPT was offered by the experimenter-employer, more than 53% of female participants opted for AI assistance, compared to 57% of males, in a gender-neutral online task. This shift prompts a reevaluation of gender gap trends in competition entry rates, particularly as women increasingly embrace generative AI tools, resulting in a boost in their confidence. We rule out differences in risk aversion. The discussion suggests that these behavioral patterns may have significant policy implications, as the introduction of generative AI tools in the workplace can be leveraged to rectify gender disparities.

Keywords: Gender differences, ChatGPT, Competition, Economic experiments

JEL classifications: C90 J16 J71

I. INTRODUCTION

Despite notable progress in reducing the gender wage gap in recent decades, gender disparities persist significantly at the upper levels of the professional hierarchy. Notably, research by Blau and Kahn (2017) underscores that, particularly in the United States, the narrowing of the gender wage gap has been more pronounced in the middle and lower segments of the income distribution, while the disparities at the upper end have proven remarkably resilient. Further emphasizing this disparity, data from the Global Gender Gap Report 2020 reveals a stark underrepresentation of women in senior management positions globally. Only 36 percent of top-level executives in both the private and public sectors are women, and a mere 18 percent of companies are led by female leaders.

In general, several factors related to the workplace may affect the gender gap including the degree to which compensation is linked to relative performance and whether the arrangement is team-based (Flory et al. 2015). Evidence also suggests that competitive incentives in the workplace are a much stronger "turn off" for women. Competitive workplaces (defined as those with an "individual tournament-based" pay approach which features a significant proportion of variable pay based on competition with other workers) significantly increase the gender gap in application, with women's likelihood of applying for the position dropping substantially relative to that of men (Buser et al. 2014).

A large literature also documents the link between the gender gap and the gender differences in competitiveness showing robust evidence that women are more reluctant to compete than men (Croson & Gneezy, 2009; Gneezy et al. 2003 and Niederle & Vesterlund, 2007). Drawing on quasi-experiments conducted in both real-world and laboratory settings, numerous studies have identified various factors influencing gender differences in the willingness to compete. One prominent factor is the tendency of males to display over-confidence, as highlighted in the work of Moore and Schatz (2017). Other contributors include the impact of nurturing, as suggested by Booth and Nolen (2012), differences in risk attitudes (van Veldhuizen, 2022), time constraints (Shurchkov, 2012), males' self-esteem (Charness et al., 2018), luck (Gill & Prowse, 2014), job uncertainty (Flory et al., 2015), time preferences (Charness et al., 2022), and the overall socialization process (Andersen et al., 2013).

Conversely, certain scenarios have been identified where women's competitiveness intentions increase. Research indicates that when females have the opportunity to select the gender of their co-participant (Datta Gupta et al., 2013) and when they find the job meaningful in terms of the task nature, women tend to demonstrate higher levels of competitiveness (Burbano et al., 2023).

In light of these insights, our investigation focuses on exploring whether the adoption of AI-tools contributes to encouraging females to express a competitive intention, aiming to mitigate the gender gap in competitiveness.

Already, Young et al. (2023) pointed out that the new generative AI technology potentially offers a rare opportunity to disrupt traditionally male-dominated fields in the labor markets and make diversity a priority early on. Yet, the lack of women in the rapidly expanding fields of AI and data science is already noticeable¹. Women should fully participate in the data science workforce and use generative AI tools for the gender gap to be rectified (Segovia-Pérez et al., 2020).

Artificial Intelligence (AI), as defined by Taddy (2018), denotes a system with the capacity to assimilate human-level knowledge, thereby expediting or automating tasks that were conventionally

¹ Women make up 32% of workers in AI and data roles worldwide (World Economic Forum, the global gender gap report, 2021), and only 18% of users across the largest online global data science platforms (Young et al., 2021).

carried out by humans. For that reason, it is growing fast with potential for far-ranging economic and societal effects. Its proponents, and now even some previously skeptical experts, believe that it will revolutionize white-collar and male-dominated work (Grace et al., 2024). Numerous firms are actively investing in AI technologies to streamline labor processes (Babina et al., 2024), reduce operating costs (Acemoglu et al., 2020), spur product innovation (Braguinsky et al., 2021), enhance their customer service (Luo et al., 2021), detecting emerging risks (Kim et al., 2023) and foster overall company growth (Babina et al., 2024). These investments are resulting in transformative changes, influencing job-entry decisions (Acemoglu et al., 2022), and altering workforce compositions in terms of gender and skills (Agrawal et al., 2019 and Babina et al., 2022). Thus, the impact of this innovative technology on workers' behavior and responses in an AI context remains unexplored (Felten et al., 2023 and Korinek, 2023).

In this paper, we undertake an experimental investigation to examine the influence of an AI-tool on the gender gap. Our experiment is structured based on the design by Niederle and Vesterlund (2007), with ChatGPT serving as our induced treatment. Acknowledging the existing utilization of ChatGPT in online labor markets, we employ a novel strategy to track participants in the control group who independently choose to use it. This enables us to ascertain the actual extent of the gender gap and subsequently explore how it manifests among participants who use it voluntarily (control group) or opt for it deliberately (treatment group).

It appears that an initial 11 p.p. gender gap diminishes, with females predominantly choosing ChatGPT showing a greater likelihood of entering the tournament compared to their male counterparts. To explore the reasons behind this trend, we investigate how confidence influences this relationship. Our analysis reveals that females display overconfidence only when they opt for the provided treatment of ChatGPT, leading to an even higher probability of engaging in the competitive environment of the tournament. Importantly, we find no evidence indicating differences in risk aversion levels between men and women attributable to the utilization or selection of the AI tool.

Our research aligns with and enriches three streams of literature. Firstly, our findings extend the economics of artificial intelligence (AI), an area receiving considerable attention in labor economics today. The impact of AI on jobs is a hotly debated topic, with some studies suggesting it may displace jobs, especially routine ones, akin to automation. On the flip side, other studies argue that AI can create new opportunities, particularly in high-skilled jobs, boosting productivity and overall economic growth (Acemoglu et al., 2021; 2022; Webb, 2019). Using economic principles and experiments, we uncover various effects of AI on workers' behavior (Roth, 2015)². Our research goes a step further by examining the personality profiles of employees adopting AI tools and how it influences their productivity in crowdsourcing microtasks and their intention to compete.

Secondly, our results highlight external factors that can help narrow the gender gap in competitive work settings. Aligning with previous studies, we identify a baseline gender difference in tournament entry, showing that this gap can be attributed to differences in confidence, particularly among female participants, shaped by the adoption of ChatGPT in the labor process. For instance, prior studies indicate that in the main treatment, men enter tournaments at twice the rate of women, but this difference disappears when considering the entrants' confidence levels (Charness et al., 2018; Markowsky & Beblo, 2022; van Veldhuizen, 2022).

² Adopting a broad definition from (Roth 2015), the economics of AI investigates the repercussions of AI on resource allocation among participants and the operative mechanisms behind it.

Thirdly, we contribute to the research on the future of work and online labor markets, with a focus on investigating how the emergence of AI technology, like ChatGPT, impacts the online labor market (Lysyakov & Viswanathan, 2023). Our study reveals that if generative AI tools are provided properly to online workers, it can significantly enhance their performance and the quality of results (Qiao et al., 2023; Bahn & Strobel, 2023).

The rest of the paper is organized as follows. Section 2 we introduce the theoretical framework and our hypotheses. Section 3 describes the experimental design. In Section 4 we present the results. Section 5 includes the discussion and Section 6 concludes.

II. THEORY AND HYPOTHESES

We employ an online experiment designed to examine the hypotheses outlined below.

Hypothesis 1: *The utilization of ChatGPT is expected to exhibit a gender disparity within the control group, with a higher proportion of males engaging with the platform compared to females. This discrepancy may stem from a perception among females that the use of ChatGPT is indicative of online misbehaviour and cheating.*

Hypothesis 1 builds on evidence suggesting a general gender disparity in AI adoption. Recent data from a 2023 survey highlights a significant gap, with 54% of men incorporating AI into their lives compared to only 35% of women (Pew Research Center, 2023)³. This divergence may stem from psychological studies indicating that women tend to require a higher level of competence before embracing new technology, while men are more open to exploring AI without extensive proficiency (Venkatesh et al., 2004 and Stoet & Geary, 2018). Moreover, women may perceive using ChatGPT independently in a task as a potential sign of misconduct or cheating. A 2020 report from the European Institute for Gender Equality indicates that only 54% of women hold positive views about AI tools, compared to 67% of men⁴. Using ChatGPT in the workplace is often viewed as a signal of employee misbehavior, leading to its prohibition by many companies like Amazon, Microsoft, and Spotify⁵ (Bin-Nashwan et al., 2023). Additionally, studies show that women are inclined to act more ethically than men (Arlow, 1991; Miesing & Preble, 1985 and Tyson, 1992), with a higher propensity for ethical behavior in the workplace and within working groups (Bowles & Gelfand, 2010; Hillebrandt & Barclay, 2020 and Chadi & Homolka, 2022). Collectively, these studies underscore the substantial gender norms disparity in perceptions of technology usage at work (Huffman et al., 2013).

³ Most Americans haven't used ChatGPT." Pew Research Center, Washington, D.C. (May, 2023).

⁴ https://eige.europa.eu/publications-resources/toolkits-guides/gender-equality-index-2020-report/gendered-patterns-use-new-technologies?language_content_entity=en

⁵ <https://jaxon.ai/list-of-companies-that-have-banned-chatgpt/>

Hypothesis 2: *In the treatment group, it is anticipated that a greater number of females will seize the opportunity to use ChatGPT compared to males.*

Hypothesis 2 posits that within the treatment group, a higher proportion of females are expected to embrace the opportunity to utilize ChatGPT compared to their male counterparts. This anticipation is grounded in two key factors. Firstly, it is hypothesized that the AI supply will function as an educational booster for females (Charness et al., 2022). Historically, there has been a gender gap in access to educational resources, and the introduction of AI technologies, such as ChatGPT, may serve as a valuable tool to bridge this gap, providing females with enhanced learning opportunities (Bao et al., 2024 and Zhang et al., 2019). In fact, previous work suggests that women tend to overinvest in educational tools in order to be more prepared in certain working settings (Chen & Chevalier, 2012 and Sinning, 2017). Giving women an external assistance may be a way to reduce the stress of competition, which has been found to play an important role in creating a gender gap in tournament entry (Shurchkov, 2012). Secondly, the hypothesis suggests that females place a higher value on seizing rewards offered by employers compared to males. In the context of the workplace, the AI supply by the employer may be perceived as a reward, and this potential discrepancy in perceived value may contribute to a greater willingness among females to engage with and leverage ChatGPT in comparison to their male counterparts. For example, Avery et al., 2023 showed that women are relatively more likely to complete their job application when they are assessed by AI tool given externally. These factors collectively form the foundation for the hypothesis that a higher number of females will choose to utilize ChatGPT within the treatment group.

Hypothesis 3: *We will find an unconditional gender gap because males enter the tournament more than females because they exhibit higher level of overconfidence and have a higher propensity to compete.*

Psychological studies consistently reveal a common trend: both men and women tend to exhibit overconfidence regarding their performance, yet research indicates that men typically display a higher degree of overconfidence compared to women (Kahneman et al., 1982; Beyer, 1990 and Beyer & Bowden, 1997). This pattern is further supported by findings from Barber and Odean (2001), who demonstrate that men engage in more excessive trading than their female counterparts in financial markets. In the context of our experiment, if men indeed demonstrate greater overconfidence in their relative performance, it is anticipated that the likelihood of choosing to participate in the competition will be higher for men than for women with equivalent performance levels (Niederle & Vesterlund, 2007). Moreover, Women might exhibit a higher reluctance to enter competitive settings due to a potential aversion to performing under competitive conditions. The anticipation of psychic costs associated with participating in future competitions may dissuade women from engaging in tournaments. Conversely, men may perceive a psychic benefit in anticipation of such scenarios, leading them to be more inclined and drawn towards competitive environments. Already, existing research has presented compelling evidence pointing to a gender disparity in participation in tournaments, even in tasks perceived as gender neutral. Notable studies, such as those conducted by Apicella and Dreber (2015) on a rope-skipping task and Apicella et al., (2017) on the counting-zeros task, have demonstrated this phenomenon. Moreover, gender gaps have been observed in certain instances, even with tasks traditionally perceived as female-typed, as illustrated by studies like

Wozniak et al., (2014) and Klinowski (2019). Testing this hypothesis is a replication exercise and our initial point of analysis.

Hypothesis 4: *Women who choose to engage with ChatGPT will experience a significant increase in self-confidence, subsequently leading to a reduction in the gender gap observed in competitive tournament entries.*

In existing literature, it has been consistently noted that women tend to exhibit lower levels of self-confidence compared to men in a wide range of competitive work environments. This discrepancy is often attributed to various external and societal factors, which impose greater "internal" or "psychological" barriers on women (Lenney, 1977 and Instone et al., 1983). The significance of self-confidence in overall well-being and its pivotal role in optimizing performance within professional settings have been underscored in studies (Compte & Postlewaite, 2004 and Koszegi, 2006). Moreover, external conditions have been identified as influential factors in shaping individuals' self-confidence (Barber & Odean, 2001). In light of this understanding, our hypothesis posits that the adoption of our generative AI tool, ChatGPT, by women will lead to an elevation in their self-confidence levels. This boost in self-confidence is anticipated to result in heightened productivity and more efficient task execution. As a direct consequence, we expect a reduction in the gender gap observed in competitiveness, as women may feel more empowered to participate and excel in competitive conditions.

III. EXPERIMENT

Economic impacts of artificial intelligence (AI)

In our experimental design, we employed ChatGPT (Chat Generative Pre-Trained Transformer), a prominent instance of Large Language Models (LLMs hereafter)⁶. These models display advanced applications of machine learning algorithms, demonstrating a generative capacity for producing original content, solidifying their position as generative Artificial Intelligence (AI) (Bubeck et al., 2023)⁷. Previous studies have already explored various mechanisms through which AI interfaces with labor markets.

First, AI could function as a direct replacement for human workers, especially those involved in routine tasks that are prone to complete automation by this technology (Autor, 2022 and Gallego & Kurer, 2022). Second, within specific occupations, AI has the potential to enhance or fine-tune human labor, consequently boosting productivity and quality, functioning as a complementary asset to human labor (Felten et al., 2021;2023). Third, AI holds the potential to create new employment opportunities, requiring human labor for the development, maintenance, or utilization of AI to accomplish tasks that were previously beyond human capabilities (Acemoglu et al., 2022). For that reason, this new technology proves to be highly effective in economics and within experimental designs, particularly serving as a key feature in online experiments (Charness et al., 2023). Our study employs an experiment conducted in an online labor market (OLM, hereafter) to investigate the impact of ChatGPT in connection to the second point mentioned above. It represents one of the initial efforts to

⁶ LLMs are a class of AI models that have a large number of parameters (175 billion for ChatGPT) and are trained on large datasets of text.

⁷ According to the Organization for Economic Co-operation and Development (2019), an AI system is defined as a “*Machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions.*”

provide evidence on how ChatGPT can contribute to the labor process in a manner that reduces the gender gap.

Already in the online labor context studies have investigated how ChatGPT is affecting workers' behavior, and the results are mixed. Lysyakov & Viswanathan, 2023, give evidence that the introduction of ChatGPT substantially raised online workers' average productivity (Noy & Zhang, 2023)⁸ and the probability to get hired (van Inwegen et al., 2023). Horton (2023) also explored the use of generative AI as simulated economic agents, and he experimentally explored their behavior. In this direction, it also seems that, generative AI tools can outperform crowd workers in several online annotation tasks (Gilardi et al., 2023) resulting in online workers' experiencing reductions in both transaction volume for online gigs (Liu et al., 2023), employment and earnings (Hui et al., 2023 and Yilmaz et al., 2023). Lastly, Awad et al., 2023 with an online experiment revealed that the utilization of AI-tools does not affect the quality and gender diversity of applicants compared to human evaluators.

By taking into consideration all the above-mentioned studies, we see that ChatGPT is by far the most popular LLM. ChatGPT was released by OpenAI in November 2022 and can accommodate an array of downstream applications (Wand et al., 2022). Immediate use cases of GPT encompass content writing, copyediting, answering questions, and language translation. Beyond its capabilities, ChatGPT is not only free to use but also accessible to the public⁹. Not surprisingly, the user base of ChatGPT has experienced rapid growth, reaching 100 million since its launch. As of the time of writing this paper, Large Language Models (LLMs) are in a phase of rapid development, with recent advancements such as GPT-4 and Bard.

Our study explores the impact of ChatGPT on workers' behavior within the context of Amazon Mechanical Turk (MTurk), one of the largest crowdsourcing online labor markets globally. OLMs are not merely a reflection of other employment contexts; they inherently represent expanding economic institutions that are progressively adopting characteristics akin to other workplaces in the post-pandemic era (Parker & Grote, 2022). They can effectively reconcile the supply and demand of labor across time and space offering flexibility as a key element in shaping the future of work (Chen et al., 2019). For that reason, these platforms present valuable contexts for studying behavioral experiments, given their dynamic and evolving nature (Horton et al., 2011). Many MTurk experiments have already been linked, workers' working behavior with their co-creation intention (Kazai et al., 2011, 2012), learning intention (Kokkodis & Ransbotham, 2023), incentives (Mourelatos et al., 2023), mood (Mourelatos, 2023) and personality traits (Mourelatos et al., 2022).

AI detection

To carry out the pre-phase step of the online experiment, we had to create a procedure that allowed us to determine the source of the text we received from users, whether it was entirely user-generated or aided by AI. For this task, we evaluated different services designed for this purpose and devised a technique to log the user's actions. This was made possible due to the unique specifications of the experiment, such as users entering text into an input box that featured methods for recording the text input history. The development of the method was inspired by research examining typing behavior

⁸ Authors demonstrate in a controlled experiment that ChatGPT made writers 40 percent faster while improving output quality and helping writers with weaker skills more, thereby reducing output inequality.

⁹ Dell'Acqua et al., 2023 suggest that the capabilities of generative AI create a "jagged technological frontier" in which some tasks can be easily done by AI, while others, though seemingly similar in difficulty level, are outside the current capability context of AI.

and the authentication of individuals in educational environments. For example, (Roth et al., 2014) studied the use of typing dynamics for continuous user authentication, while (Leinonen et al., 2016) examined the identification of students in programming exams using typing patterns, highlighting the potential of typing behavior for verifying the authenticity of user submissions.

Our AI-detection system involves the development of a rule-driven system that can identify if users have used ChatGPT to generate their answers. This system carefully records and analyzes user interactions to determine the origin of the content. In the pre-phase step on the experiment page, the system diligently captures extensive data on user interactions. This involves monitoring typed keys, instances of copied content, and submitted text. Moreover, the monitoring also includes tracking copy events like copying the question text and recording the time it happened, detecting tab changes by the user, and identifying when a participant clicks the button (particularly for the GPT group, which redirects them to the GPT page). To preserve the sequence of events, each action is timestamped. The analysis focuses on the frequency and timing of keystrokes to detect patterns in manual typing.

Simultaneously, the system closely examines paste events, particularly when large blocks of text are inserted at once. These events indicate the use of generative AI tools like ChatGPT to source content from external platforms. Moreover, the approach involves monitoring browser tab changes to understand how users behave, as they may refer to external sources or utilize AI tools while crafting their responses.

Using the collected data and analysis, the system classifies submissions into two main categories: (a) manually typed responses (user-created content) and (b) those that potentially utilize ChatGPT for content creation (ChatGPT-generated content). Metrics such as the total keypress count, paste events, and average time interval between keystrokes provide additional support for this classification, ensuring a comprehensive evaluation. Furthermore, the system acknowledges that users may choose a hybrid approach for content creation, combining manual typing with pasting content from external sources. Cases demonstrating this combination are labeled as ChatGPT-generated content¹⁰.

Building upon the methodology described, we also explored the potential of utilizing online services that can detect AI-generated text. However, according to existing research, the capacity of machine learning tools to accurately detect AI-generated text is restricted (Anderson, et al., 2023). According to Perkins (2023), the text generated by LLMs can frequently resemble human-authored content, which presents considerable difficulties in accurately identifying such texts. In a related study, El-Sayed et al. (2022) achieved only a 59.5% accuracy in differentiating between text from humans and language models like ChatGPT. To verify the aforementioned findings, we conducted a series of evaluations using various tools. These verifications reinforced the previously mentioned points. Beyond the problems identified, an additional issue in our use case was the length of the responses from the workers, which for reasons of avoiding fatigue, had an upper limit of 150 words.

This disparity in effectiveness emphasizes the challenge of relying exclusively on textual analysis for AI detection, especially in instances involving shorter texts or in contexts where AI-generated content closely imitates human writing styles. On the other hand, our rule-based system makes use of direct interaction data (keystrokes, paste events, tab changes) which are less prone to the restrictions imposed on text analysis algorithms.

¹⁰ For example, assume a situation where a worker starts editing their answer. Our AI-detection approach records keystrokes, highlighting a human-like typing rhythm with natural variances in speed and real-time corrections (no paste events; presence of 'Delete' and 'Backspace'). The lack of paste events implies that the content was not copied from chatGPT. When the text is finally submitted, it perfectly aligns with the keystroke log, confirming that the user manually typed the content (consistent final text with keystroke log). The case at hand serves as an example of content that was manually crafted, showcasing the user's original input without any support from ChatGPT.

Online Labor Market and Online Job

We recruited participants from the USA through Amazon Mechanical Turk, an online labour market (MTurk hereafter). MTurk involves two major categories of participants. In the context of bibliography, requesters are employers who utilize online platforms to advertise job opportunities. On the other hand, individuals known as “Turkers” participate in task-based labor and receive compensation as set by the requesters. Requesters can hire Turkers based on their HIT experience or task completion approval rate. The workflow starts with a requester posting the main task as an open call. Turkers choose HITs according to their preferences. Subsequently, the requesters proceed to evaluate the work submitted by each participant, determining its acceptance or rejection based on its alignment with the job requirements established at the commencement of the task. Eventually, MTurk ensures that the payment posted by the requester is distributed to the approved Turkers. Thus, MTurk’s crowdsourcing flow was very compatible with our experiment’s workflow and characteristics, covering the basic components of our online task characteristics. There have been several studies indicating that data gathered through AMT is as reliable as data collected in a traditional physical lab (Arechar et al., 2018).

Following an experimenter-as-employer paradigm (Horton et al. 2011) and drawing on Niederle & Vesterlund, 2007’s experimental framework our experiment involves participants solving a task in two different schemes: noncompetitive piece-rate and competitive tournament. Participants are subsequently instructed to select the compensation scheme they desire to apply to their upcoming performance. By engaging in this, participants can acquire firsthand experience with both compensation forms, enabling us to evaluate whether individuals of equal performance select the same compensation scheme. The aim of our experiment entails the summation of sets of two two-digit numbers¹¹. The numbers are drawn randomly, and each problem is presented using the following format, wherein participants are required to fill in the sum in the blank box (Figure 1).

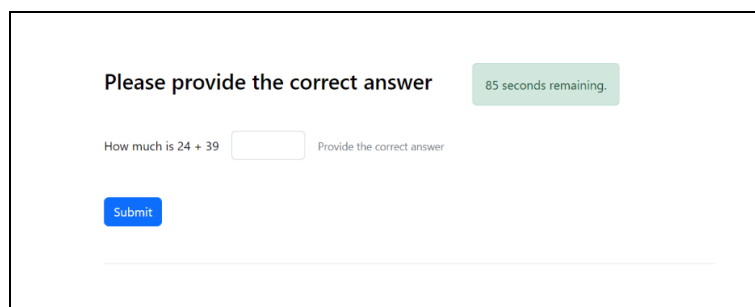


Figure 1. Sample addition task case.

When participants enter their response on the task webpage, they receive a new problem without receiving any feedback on the correctness of their previous answer, and they do not have access to a record of the number of correct and incorrect answers. Participants have 90 seconds in each experimental phase in which they may solve as many problems as they can. Time expired automatically. Participants cannot advance to the next page before the 90 s elapse. The final score is

¹¹ We have selected a simplified version involving the addition of two two-digit numbers, in contrast to the five-digit addition utilized in Niederle and Vesterlund’s 2007 experiment. This choice is made to mitigate the potential for cheating effects and misbehavior, as highlighted by List and Momeni in 2020. Additionally, it aims to minimize any potential boredom and fatigue effects that may arise, particularly given the online nature of the experiment (Zheng et al., in 2011).

determined by the number of correctly solved problems. We opt for this task for three key reasons: (1) it effectively mitigates the chances of cheating in the online experiment setting, (2) past research has demonstrated a gender gap in tournament entry within a baseline condition closely resembling ours with this task, and (3) it is both perceived and observed to be gender-neutral. This will enable us to rule out performance differences as an explanation for gender differences in tournament entry.

Experimental Design

The experiment comprised three distinct phases: an initial survey, a pre-task segment, and the primary set of tasks. We employed the survey to control for demographic, socioeconomic characteristics, and psychological gender differences. The pre-task segment was designed to identify the use of ChatGPT, while the set of tasks aimed to measure gender differences in competitiveness.

At the commencement of the experimental session, participants were required to fill out a survey questionnaire having information on demographics (including gender, marital status, age, ethnicity, and educational levels), socio-economic variables (such as the FAS index, health status, income, worker status, and primary source of income), personality traits (big five personality traits), and other psychological-related variables (encompassing motivation for participation, self-esteem, and the honest humility variables). For the aforementioned questionnaire, we utilized a short 10-item, five-point Likert Scale to measure personality traits, drawing from the Big Five personality test (John & Srivastava, 1999 and Costa & McCrae, 1999). This resulted in the derivation of our five personality variables: openness, conscientiousness, extraversion, agreeableness, and neuroticism. Self-esteem was assessed using a single construct based on Rosenberg's scale (Rosenberg, 1965 and Robins et al., 2001). Honest humility was gauged through a 10-item, five-point Likert Scale (Ashton et al., 2014 and Hilbig et al., 2014), resulting in two variables: fairness and modesty. The Family Affluence Scale (FAS index), measured on a six-item Likert scale, was employed to assess socioeconomic status. This inventory, recognized for its validity and ease of use, is widely accepted and has been utilized in various studies measuring wealth, including studies such as Boyce et al. (2006).

Before engaging in the primary set of tasks, participants went through a second phase, which involved providing a free-text answer to the question: "Can you share your profession without directly naming it? Tips: (a) Use synonyms or alternative words that describe your profession without giving it away directly; and (b) Mention examples of tasks you perform or tools you utilize, instead of stating your job title." We utilized this question for two purposes. Our initial aim was to include a personal element in the response, enhancing recognition through either manual inspection or the use of automatic approaches. Additionally, to enhance the level of difficulty and stimulate the participant's use of ChatGPT to complete the task. The limit of 150 words was established with the intention of marginally enabling the utilization of automated techniques for identifying AI-generated text, while also preventing his fatigue and demotivation. All participants were randomly assigned to either the control or treatment group for this pre-task, employing a uniform distribution algorithm. Control group participants were tasked with completing the pre-task independently on our webpage, without external assistance. In contrast, the treatment group had the added advantage of utilizing a ChatGPT-like AI tool. This tool was conveniently accessible through a button embedded in the webpage, redirecting them to an external page where they could seek assistance in formulating their answers. This tool remained available and active throughout the experimental process. Ultimately, the text submitted by both groups underwent analysis using the AI-detection tool to assess adherence to instructions and determine whether participants in either group had employed ChatGPT.

Upon completion, all participants proceeded to engage in our experiment. Specifically, as illustrated in Figure 2, participants were required to complete four tasks:

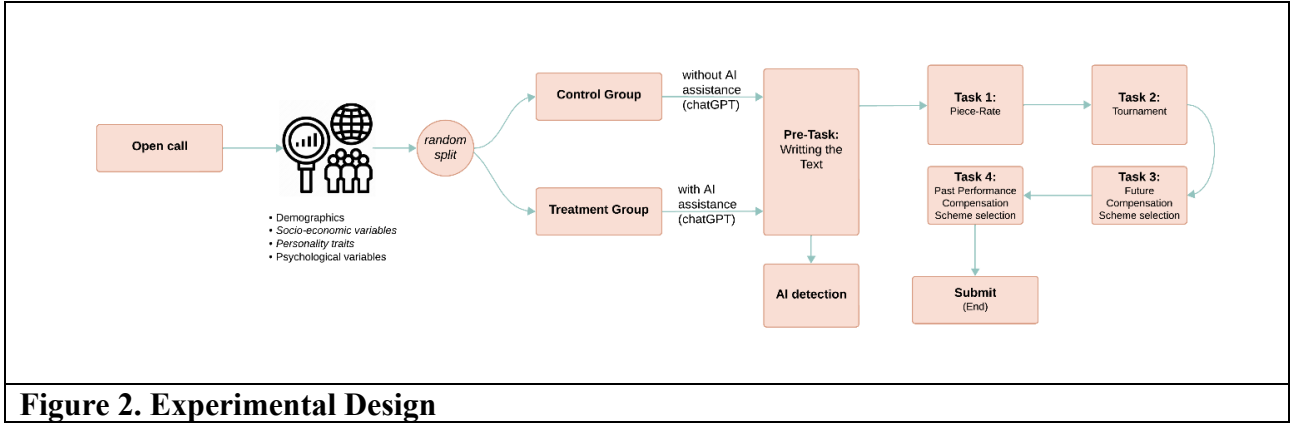
Task 1—Piece Rate compensation scheme: participants assigned the online job of the summation of several sets of two two-digit numbers and evaluated based on a piece rate system. If Task-1 is randomly selected for payment, a sum of 25 cents will be given for every correct response.

Task 2—Tournament compensation scheme: participants are randomly assigned to groups of four and they are required to complete again the online job of the summation of several sets of two two-digit numbers. If Task-2 is chosen at random for payment, the participant who solves the most correct problems in the group will receive \$1 for each accurate response, while the remaining participants will not receive payment. In case of tied scores, the highest scorers will be chosen randomly to decide the winner. The tournament is designed so that participants with a 25 percent chance of winning receive the same expected payoff as the piece-rate compensation scheme.

Task 3—Selection of Compensation Scheme for Future Performance: Before commencing their third round of online tasks, participants make a critical decision regarding their compensation method. They can opt for a piece-rate of 25 cents per correct answer like Task-1 or choose the tournament option like Task-2. In the tournament scenario, participants earn \$1 per correct answer if their Task-3 score surpasses those achieved in the Task-2 tournament; otherwise, they receive no compensation. In the case of ties, a random winner is determined. This approach ensures that the competition is based on prior performances under similar conditions, reduces errors stemming from biased beliefs about others' choices, and eliminates the influence of one's choice on others' earnings. Consequently, the decision to compete relies solely on one's ability to outperform the Task-2 scores of others and their preference for tournament-based competition.

Task 4—Choice of Compensation Scheme for Past Piece-Rate Performance: To discern the underlying reasons for the gender gap in tournament entry, specifically whether it is driven by gender-based preferences for competitive environments or can be attributed to broader factors like risk aversion, we present participants with a concluding task. In this scenario, participants are presented with a choice akin to Task-3, but without the requirement of using a tournament performance and participating in a subsequent tournament. Also, participants are not required to perform this task; instead, if selected for payment, their compensation is based on their Task-1 piece-rate correct answers. They choose between a piece-rate of 25 cent rate per correct answer or a tournament for their past performance. In the tournament, winning means earning \$1 per correct answer if their Task-1 performance tops their group; otherwise, they get nothing, with ties resolved randomly. Before choosing, they are reminded of their Task-1 performance.

Lastly, we also gather participants' beliefs about their relative past performance to link these perceptions to their compensation choice, especially examining if gender-based confidence disparities contribute to tournament entry decisions. This assessment helps explore the role of overconfidence in mediating the gender gap in competition entries. Thus, at the end of the experiment participants are asked to guess their rank on the Task-2 tournament. Each participant picks a rank between 1 and 4 and is paid \$0.50 if the guess is correct.



To assure data quality, we set 2 default criteria for workers to have the opportunity to be hired in the experiment: 80% success rate in their previous task completion activity and participation in at least 50 approved tasks in MTurk (Peer et al. 2014). We collected data from 1233 participants. To avoid self-selection bias, the offered wage was in line with MTurk price policy. Upon conclusion of the experiment, a random selection is made from numbers one to four, determining the specific task for which participants will receive earnings. The experiment, spanning approximately 10 minutes, resulted in average earnings of \$9.20 for participants.

Pilot Studies

In preparation for our main experiment, we conducted two preliminary pilot studies to measure the accuracy of our AI detection strategy and verify the experimental sequence. The initial pilot study involved 98 students from the University of Peloponnese. We tasked students in the class exclusively undertaking the second phase of our experiment, focusing on text generation. Through random assignment, half of the students produced the text without AI assistance, while the remaining half utilized AI assistance, specifically from ChatGPT. This methodology ensured that we had a reliable ground truth for each case. Intriguingly, our AI detection strategy accurately identified all instances where texts were generated without AI assistance, while achieving a detection rate of 49 out of 50 for the texts generated with ChatGPT (98% efficiency).

In addition to the pilot study aimed at assessing the effectiveness of our AI-generated text detection method, we conducted a separate pilot experiment on Amazon Mechanical Turk (MTurk). The primary objective of this experiment was to scrutinize our experimental procedure for any potential design flaws. A total of 102 users actively participated in this pilot. Overall, the execution of the pilot proceeded smoothly. However, we encountered a few instances where users managed to bypass our geographical restrictions by employing a Virtual Private Network (VPN) from countries such as India, Nepal, and Bangladesh. To address this challenge and ensure the inclusion of only U.S. citizens in our main experiment, we implemented a service capable of detecting VPN usage. To achieve this, we utilized Cloudflare's technology¹², which not only enhances website security and performance, but also enables us to determine the country a user is from by analyzing their internet connection. This technology even has the capability to detect attempts to conceal one's location using the Tor network, a commonly employed method for hiding a user's whereabouts. Furthermore, to address the issue of users bypassing our location requirements with VPNs, we employed the services of VPNapi.io¹³. This

¹² <https://en.wikipedia.org/wiki/Cloudflare>

¹³ <https://vpnapi.io>

service verifies each user's IP address, which serves as a unique identifier for their internet connection, against a comprehensive database to determine if it originates from a VPN. This step was crucial in maintaining the fairness of our experiment and ensuring that only individuals genuinely located in the U.S. were included. By combining these tools, we took the necessary precautions to ensure that all participants in our study were truly U.S. citizens, thus upholding the integrity of our experiment and ensuring the validity and reliability of the results. Users identified using a VPN were subsequently excluded from participating in the experiment.

Next, based on the pilot results, a power analysis was conducted to identify a minimum necessary survey sample and confirm the validity of our findings. Due to the nature of this study, which examines working behaviours and responses within an online community (i.e., within the MTurk platform), we assumed that the effects on tournament entry might be small (Di Gangi et al. 2022). Therefore, in the sample size calculations, we assume that the sample would have 95% reliability about population and a sampling error of 5%. The calculations show that a threshold of a sample size of $N = 465$ in each group is required for the exact approach to attain the target power 0.9, with a significance level = 0.05. For the implementation, we use the Stata modules `samsi_reg` (Mander, 2005) and `powercal` (Newson, 2004). We excluded the responses of a few subjects who failed attention checks and those who took the survey more than once. Additionally, because subjects were randomly assigned to one of the treatments, the final distribution of subjects across treatments is not exact.

Limitations

Our study has certain limitations. Firstly, our research uncovered that a considerable number of participants in the control group independently chose to use ChatGPT. While our approach enables us to identify them, it introduces uncertainties regarding the actual percentage of participants who would opt for our treatment. This implies that within the percentage of individuals selecting our treatment, some would use ChatGPT regardless. Despite our attempts to restrict ChatGPT choice in both groups and allow only selection, we were unable to devise an effective method that wouldn't introduce noise into our experiment. Secondly, while our AI detection exhibited high efficiency in identifying participants who opted for ChatGPT during the pre-task text generation, it proved useless in detecting with accuracy individuals utilizing ChatGPT for the summation of sets of two two-digit numbers in an online job context¹⁴. Despite this limitation, we leveraged components of our AI detection strategy to monitor behavioral patterns. Notably, through the analysis of keyboard dynamics, such as copy-paste actions, we observed a significant behavioral consistency among individuals initially flagged by our AI detection algorithm. This consistency extended to the use of AI in both text-related tasks and numerical summation, as well as the reverse scenario. Third, in any experiment that permits participants to withdraw before completion, attrition presents potential challenges to the credibility of data analysis within the subset of subjects who return (refer to Hauser et al., 2019, for a comprehensive discussion, particularly in the context of online experiments). We acknowledge the significance of this issue and we tried to address it in this paper¹⁵. Thus, we conducted a comparative analysis of the demographic and behavioral characteristics of returning and

¹⁴ This stemmed from the fact that our primary AI detection strategy was initially tailored exclusively for text input.

¹⁵ We had a very small number of participants that dropped out from the experimental process ($N = 17$ participants).

non-returning subjects, finding no statistically significant differences between them, nor a systematic correlation with the treatment.

Sample characteristics, balance tests and gender differences.

Overall, the demographic profile of our sample participants closely mirrors the national distribution in the USA. Any variations observed can be attributed to the composition of online participants from Amazon Mechanical Turk, which tends to lean towards higher percentage of male participants and exhibits a deviation from the racial minority status (Ipeirotis, 2010 and US Census Bureau in 2022). Furthermore, our analysis reveals no statistically significant variations in demographics, socio-economic factors, personality traits, and other psychological variables between the control and treatment groups, as detailed in Table 1. Upon scrutinizing gender disparities within our sample, we observe a noteworthy trend—men are markedly more inclined to engage in online labor markets, even as they maintain full-time positions in the traditional labor market. However, no statistically significant differences emerge based on gender when assessing whether online labor markets serve as the primary source of income or in terms of the incentives driving participation. Moreover, our investigation unveils noteworthy gender disparities in certain personality traits and other psychological factors. Specifically, females demonstrate elevated levels of fairness and modesty, with a statistically significant p-value of less than 0.001. Employing the Big Five model for personality assessment, our analysis indicates significant gender differences. Females exhibit higher levels of openness (p-value = 0.000), while males surpass females in extraversion (p-value < 0.05), and females again surpass males in agreeableness (p-value < 0.10). Notably, although females tend to have higher levels of neuroticism, this difference does not reach statistical significance. These findings align with existing psychological literature (Costa et al. 2001). We observe no statistically significant difference in self-esteem levels between men and women.

Table 1. Sample and Balance T-tests

	Mean	Control [1]	Treatment [2]	Difference [1] – [2] [3]
<i>Demographics</i>				
Females (0/1)	0.401	0.389	0.410	-0.021
Singles (0/1)	0.152	0.155	0.148	0.007
Age	35.9	35.6	36.2	-0.56
Whites (0/1)	0.902	0.914	0.889	0.025
At least university education (0/1)	0.307	0.309	0.305	0.004
<i>Socio-economic variables</i>				
Fas index	2.119	2.131	2.106	0.025
Good health (0/1)	0.837	0.839	0.836	0.003
High income (0/1)	0.236	0.231	0.241	-0.010
Fulltime worker (0/1)	0.890	0.883	0.897	-0.014
Primary source of income (0/1)	0.224	0.222	0.226	-0.004
<i>Personality traits</i>				
Openness	6.546	6.557	6.535	0.022
Conscientiousness	6.926	6.932	6.920	0.012
Extraversion	5.734	5.678	5.790	-0.112
Agreeableness	6.564	6.603	6.525	0.078
Neuroticism	5.469	5.447	5.492	-0.045

	Psychological variables			
Extrinsic (0/1)	0.688	0.689	0.686	0.003
Self-esteem	4.743	4.816	4.669	0.147
Fairness	9.058	9.104	9.011	0.093
Modesty	4.449	4.463	4.433	0.030
Observations		610	606	

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment. High income refers to annual household income of at least 80,000 U.S. dollars.

Notes: N = 1216. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

IV. RESULTS

Who Utilizes vs Who Selects ChatGPT?

As an initial step in our analysis, we aim to explore gender-based preferences in Artificial Intelligence (AI). Two key questions in the initial questionnaire focused on AI awareness and daily usage of ChatGPT. Utilizing a two-sided t-test, we found that the difference in AI awareness between genders is not statistically significant ($p = 0.882$) (Men: 89.6% vs. Women: 89.9%). However, when assessing ChatGPT usage, a statistically significant difference emerged ($p = 0.000$) (Men: 79% vs. Women: 54.9%). This suggests that while both genders exhibit awareness of artificial intelligence, females tend to use generative AI tools, such as ChatGPT, less frequently than males.

Moving forward in our analysis, we shift our focus to investigate the factors influencing individuals' choices regarding the adoption of the ChatGPT AI tool. Recall, our study involves two distinct experimental groups: the control group and the treatment group. In the control group, ChatGPT is not provided, but we monitor for any instances where individuals independently choose to use it. Conversely, in the treatment group, we actively present individuals with the option to select ChatGPT, thereby introducing an AI supply treatment. In this stage of our analysis, we categorize our examination by groups to gain a clearer understanding of behavioral patterns associated with the use of ChatGPT in each case. In Figure 3, the utilization of ChatGPT is depicted based on gender and group categorization. Within the control group, a statistically significant difference is observed, with men constituting 57.7% and women 45.1% ($p < 0.05$). Conversely, in the treatment group, the usage rates show no statistical significance, with men at 57.5% and women at 57.4% ($p = 0.974$).

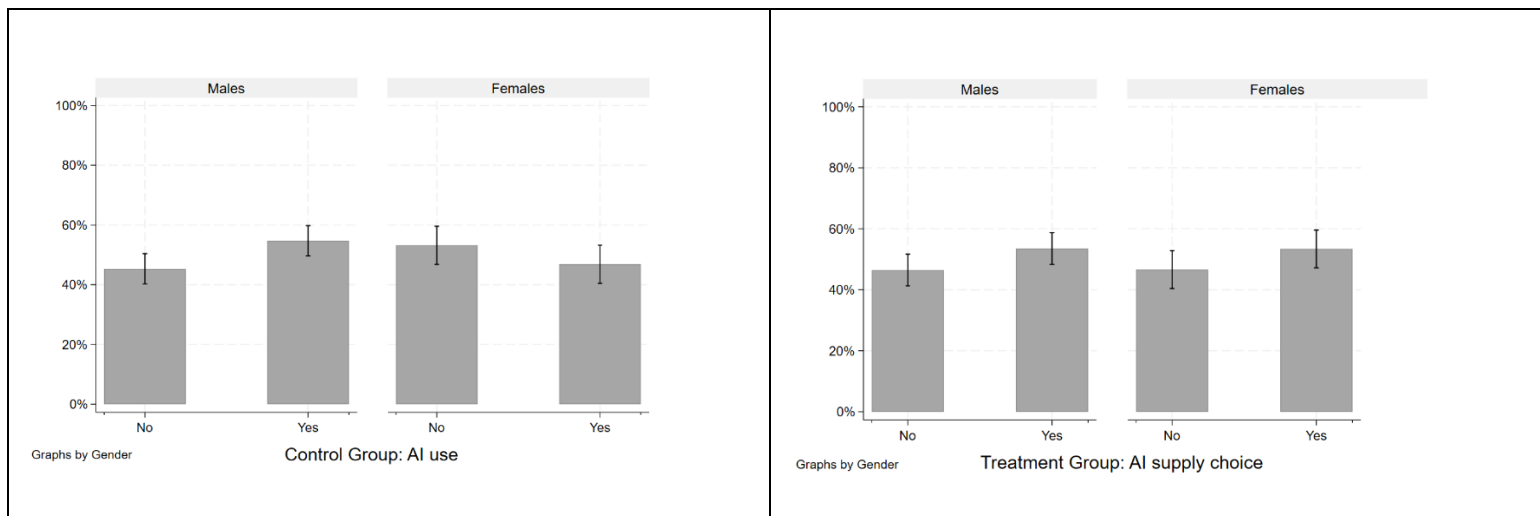


Figure 3. ChatGPT utilization based on gender and group categorization

Table 2 outlines the marginal effects on ChatGPT choice within each group. In the context of the control group, our findings indicate that females exhibit a lower likelihood, by 13.3 percentage points of opting for ChatGPT compared to their male counterparts. This pattern persists consistently across various specifications (1-4), incorporating personality traits, psychological factors, and demographic and socio-economic variables. Notably, this gender-based effect diminishes within the treatment group.

One plausible interpretation revolves around divergent perceptions of AI tools. To probe this, we gauge perceived competence for technology by posing the question, “*Do you feel that you need a high level of competence before embracing new technology? (Yes/No)*”. Simultaneously, we assess perceived misbehavior in the workplace by asking, “*Is the utilization of AI tools in the workplace indicative of misconduct or misbehavior in your opinion? (Yes/No)*”. Importantly, these inquiries are made at the experiment's conclusion to minimize awareness bias in the experimental process. Figure 4 distinctly illustrates that female participants tend to associate the use of AI tools with the need for competence and perceive it as signaling misbehavior when employed in the workplace without the employer’s knowledge.

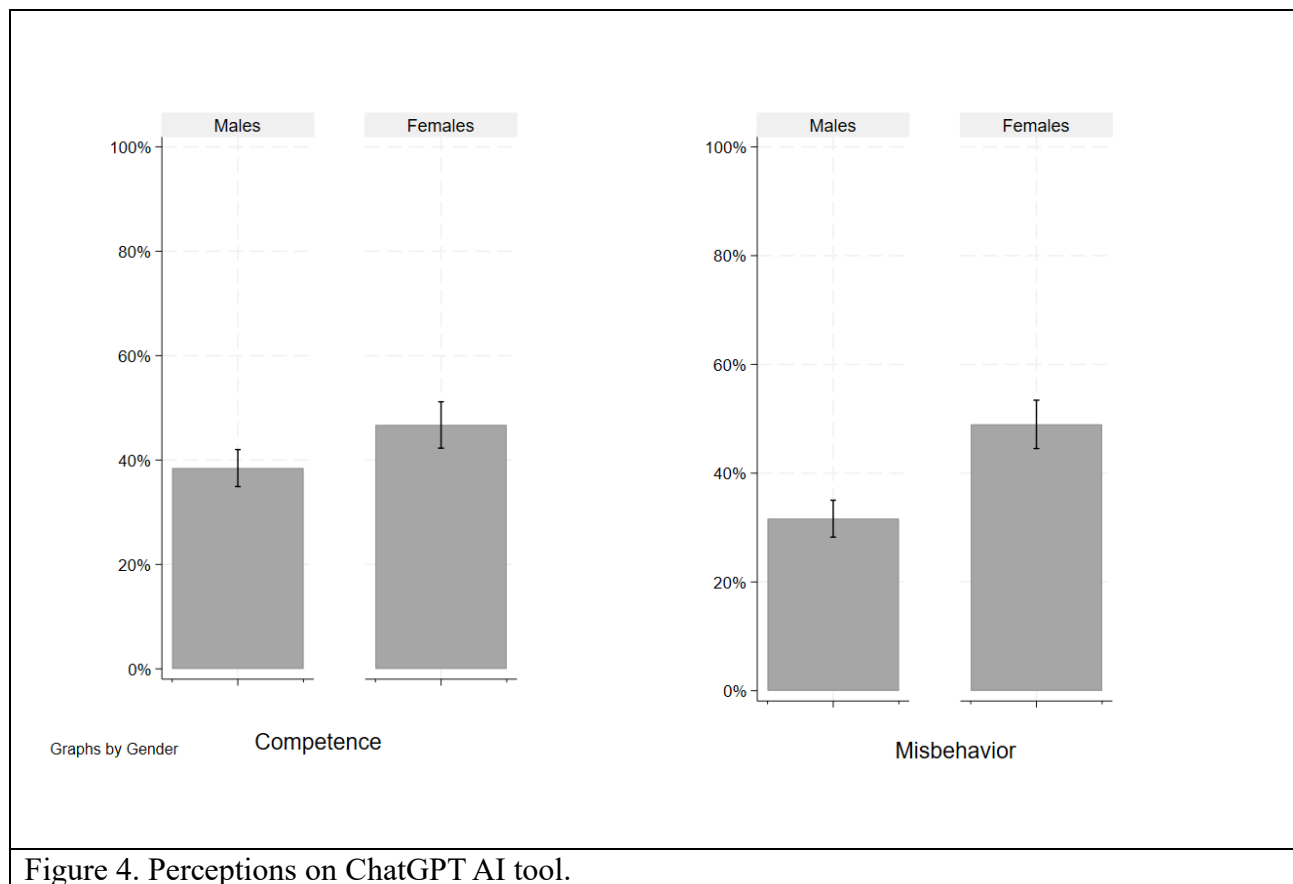


Figure 4. Perceptions on ChatGPT AI tool.

This interpretation is further substantiated by the impact of conscientiousness on ChatGPT usage within each scenario. Intriguingly, while conscientiousness exhibits a consistently negative influence on ChatGPT utilization in the control group, this effect dissipates when ChatGPT becomes an option presented by the employer-experimenter. Psychological literature unequivocally highlights

conscientiousness as a robust predictor of misbehavior and dishonesty. Higher scores in conscientiousness are associated with a diminished propensity to engage in deceitful behavior (Sackett & Wanek, 1996; Giluk & Postlethwaite, 2015; Apostolou & Panayiotou, 2019).

We note similar patterns concerning the fairness variable, assessed through the honesty-humility questionnaire (Johnson et al., 2011). Notably, the modesty variable exhibits a consistently negative impact on ChatGPT usage across both groups. This may be attributed to the notion that modesty captures aspects of individuals' beliefs that, in general, competence is a prerequisite before embracing new technology (Hilbig et al., 2014).

Thus, in accordance with our hypotheses 1 and 2, females in the control group use ChatGPT less frequently, considering it may be indicative of online misbehavior and cheating. However, when the opportunity to use ChatGPT is presented, females seize the chance more readily.

Concerning the remaining personality traits, openness exhibits a discernible impact on ChatGPT usage, particularly within the treatment group. This observation aligns with existing literature, which indicates that heightened levels of openness tend to result in a more significant inclination towards artistic pursuits, coupled with a comparatively lesser increase in inventiveness. This phenomenon is evident when considering the various facets of openness, such as fantasy, ideas, actions, as highlighted by Cubel et al. (2016). Additionally, the correlation between openness and creativity, as established by McCrae (1987), further accentuates this connection. Therefore, the variations in inventiveness and preferences for aesthetic and artistic experiences could potentially account for the divergent influence of openness on AI engagement within the context of our experiment. Concerning neuroticism, its association with ChatGPT diverges between the control and treatment groups, being negative in the former and positive in the latter. Existing literature supports the notion that individuals with elevated neuroticism levels exhibit a diminished inclination towards adopting new technological tools (Marciano et al., 2020). In the treatment group, although a direct test to interpret the positive effect is lacking, we posit that this outcome is linked to specific facets of the neuroticism trait. The facets of neuroticism, encompassing anxiety, self-consciousness, impulsiveness, and vulnerability, shed light on this phenomenon. According to personality theories, individuals scoring high in neuroticism tend to grapple with heightened emotional instability, anxiety, and negative emotions. In the context of opting for ChatGPT for work-related tasks, several reasons may elucidate this preference. For instance, the concept of "reduced pressure" suggests that ChatGPT provides a virtual, non-judgmental environment for individuals to interact. Moreover, the idea of "reduced emotional load" posits that communicating with a machine might be less emotionally demanding than interacting with humans (Gunthert et al. 1999 and Schneider, 2004).

Finally, concerning the remaining variables, consistent patterns emerge for both groups. Individuals with extrinsic incentives, higher self-esteem levels, at least tertiary education, and those engaged as full-time workers exhibit a heightened propensity to choose ChatGPT in general.

Table 2. Who Utilizes vs Who Selects ChatGPT? Probit estimates (marginal effects)

	Panel A: ChatGPT utilization = 1 (Control Group)				Panel B: ChatGPT Supply selection = 1 (Treatment Group)			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Females	-0.133*** (0.037)	-0.122*** (0.038)	-0.102*** (0.038)	-0.082** (0.039)	-0.010 (0.041)	-0.028 (0.037)	-0.028 (0.037)	0.006 (0.038)
Openness		-0.027 (0.024)	-0.018 (0.024)	-0.019 (0.024)		-0.076*** (0.023)	-0.068*** (0.023)	-0.059*** (0.023)
Conscientiousness		-0.104*** (0.020)	-0.088*** (0.021)	- 0.082*** (0.022)		-0.007 (0.019)	-0.003 (0.019)	-0.001 (0.019)
Extraversion		0.043** (0.022)	0.028 (0.022)	0.013 (0.021)		-0.009 (0.023)	-0.019 (0.022)	-0.034 (0.022)
Agreeableness		-0.032* (0.019)	-0.012 (0.020)	-0.010 (0.019)		-0.013 (0.019)	-0.006 (0.018)	-0.002 (0.019)
Neuroticism		-0.042** (0.020)	-0.027* (0.021)	-0.011 (0.021)		0.039** (0.020)	0.048** (0.020)	0.062*** (0.021)
Extrinsic			0.053 (0.073)	0.078** (0.039)			0.072** (0.038)	0.079** (0.040)
Self-esteem			0.024** (0.012)	0.029** (0.012)			0.012 (0.012)	0.018* (0.012)
Fairness			-0.059*** (0.022)	-0.053*** (0.021)			-0.005 (0.024)	-0.004 (0.023)
Modesty			-0.102*** (0.023)	-0.071*** (0.020)			-0.091*** (0.025)	-0.049** (0.021)
Age				-0.001 (0.002)				0.001 (0.002)
Singles				-0.154** (0.065)				-0.220*** (0.065)
Whites				-0.030 (0.073)				-0.066 (0.059)
At least tertiary education				0.147***				-0.104***

				(0.036)				(0.041)
Fas index				-0.183***				-0.131**
				(0.054)				(0.059)
Good health				0.093*				-0.034
				(0.056)				(0.048)
High income				-0.129***				-0.084*
				(0.048)				(0.046)
Fulltime worker				0.205***				0.299***
				(0.066)				(0.072)
Primary source of income				-0.042				0.034***
				(0.048)				(0.045)
Pseudo R ²	0.018	0.099	0.178	0.251	0.010	0.040	0.080	0.145
Obs		610					606	

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: N = 1216. Dependent variable: ChatGPT usage detection (1: Yes, 0: No).

*** p<0.01, ** p<0.05, *p<0.10.

Performance in the Online job

In both the piece-rate (Task-1) and tournament (Task-2) compensation schemes, we do not observe statistically significant differences in performance based on gender, affirming the gender-neutral nature of the online job (Niederle & Vesterlund, 2007, and Carlsson et al., 2020). For the piece-rate performance utilizing a two-sided t-test, the observed difference is not statistically significant ($p = 0.549$). For the tournament performance applying a two-sided t-test, the identified difference is also not statistically significant ($p = 0.196$). The cumulative distributions for the number of correct answers in the piece-rate (Task-1) and the tournament (Task-2) are depicted in the left and right panel of Figures 4, respectively. The performance distributions exhibit striking similarities across both genders.

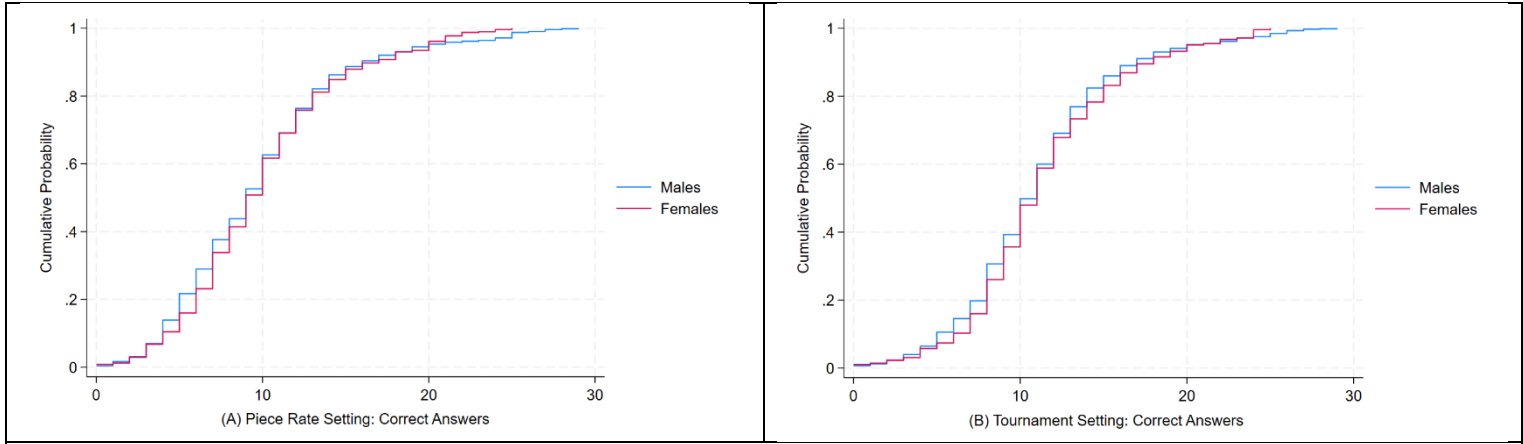


Figure 4. Cumulative distribution of performance (number of correctly solved problems) based on gender under a piece-rate compensation scheme. The left panel pertains to the entire sample, while the right panel focuses exclusively on the treatment group.

Moreover, we have observed a strong correlation in both males and females, with Pearson coefficients of 0.84 and 0.85, respectively (Figure 5). Notably, within the treatment group, females who opted for ChatGPT supply exhibited a particularly high correlation of 0.91, while those who did not choose ChatGPT supply showed a correlation of 0.83. For males, correlations were 0.86 and 0.81 for those who chose and did not choose ChatGPT supply, respectively. This indicates consistent high performance, especially among females who chose ChatGPT supply, in both Task-1 and Task-2. Furthermore, it is noteworthy that both genders generally performed significantly better under the tournament setting compared to the piece rate setting (one-sided p -value = 0.000). Our findings indicate that females showed a greater improvement in performance compared to males (1.42 correct answers for females versus 1.24 correct answers for males). However, it's important to note that this difference did not reach statistical significance. This observed trend may be attributed to learning effects or differences in incentives associated with the tournament mode of the experiment, as discussed in studies by Niederle & Vesterlund (2007) and Croson & Gneezy (2009).

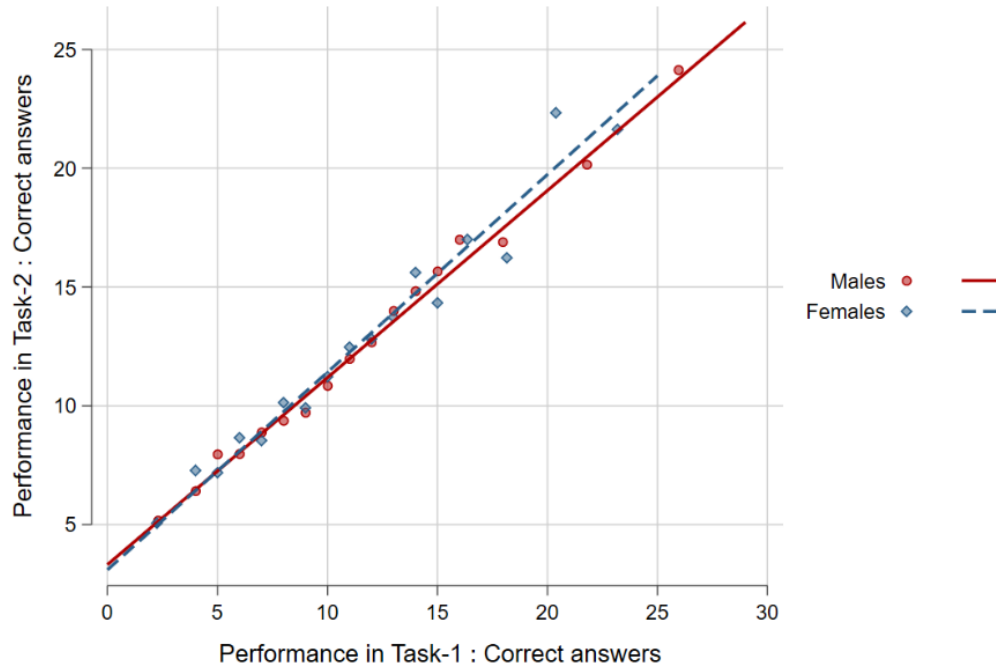


Figure 5. The relationship between performance on Task-1 and Task-2 by gender.

Analyzing Gender Disparities in Tournament Entry (Task-3 choice)

Despite comparable performance between women and men, there exists a divergence in their preferences for compensation schemes depending on whether they utilize AI or not. To explore deeper this behavioral distinction, we initially divided the sample into two groups: participants who did not utilize or select ChatGPT and those who either utilized or selected it. Table 3 displays the competitiveness choice results for participants across both groups who did not utilize ChatGPT. Probit regressions were conducted to examine the impact of participants' performance in Task-1 (piece-rate scheme) and Task-2 (tournament scheme) on their decision in Task-3 to opt for the tournament payment scheme. The analysis reveals that, while performance in the aforementioned tasks does not significantly influence the choice, gender emerges as a significant factor. Specifically, females exhibit an 11.1 percentage point lower probability of entering the competitive environment of a tournament compared to males. This finding aligns with prior research exploring the gender gap in competitiveness, as documented in various studies (Niederle & Vesterlund, 2007; Buser et al., 2021; Charness et al., 2022; Carlsson et al., 2020). To uncover potential explanations for this gender gap, we introduced a set of variables in each column (3-5). Notably, the gender gap persists consistently across all specifications, indicating its robustness and highlighting the need for further investigation into the factors contributing to this observed difference in competitiveness choices.

Table 3. Is there a gender gap when ChatGPT is not utilized or selected? Probit estimates (marginal effects)

	[1]	[2]	[3]	[4]	[5]
Female	-0.111*** (0.043)	-0.115*** (0.042)	-0.103*** (0.044)	-0.091** (0.042)	-0.084** (0.041)
Tournament performance		0.011* (0.005)	0.010* (0.005)	0.011 (0.004)	0.011 (0.005)
Tournament- piece rate change in performance		0.013 (0.008)	0.013 (0.008)	0.016* (0.008)	0.015* (0.008)
Personality Traits			✓	✓	✓
Demographic and social economic variables				✓	✓
Psychological variables					✓
Pseudo R ²	0.020	0.036	0.043	0.099	0.112

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: N = 575. Dependent variable: Task-3 choice of tournament compensation scheme (1: tournament, 0: piece-rate). Tournament performance refers to Task-2 performance and tournament- piece rate change in performance refers to the change in performance between Task-2 and Task-1. The specifications control the experimental group and incentives (extrinsic/intrinsic).

*** p<0.01, ** p<0.05, *p<0.10.

Table 4 is presented in a comparable way, illustrating the outcomes of competitiveness choices among participants in both groups who actively engaged with ChatGPT. The findings indicate that, whether individuals opted for ChatGPT through our treatment AI supply or independently, there is no discernible gender gap in the probability of entering the tournament. Notably, the shift in performance from Task-1 to Task-2 demonstrates a positive influence on tournament entry. However, it is intriguing to observe that the performance on Task-2 (tournament) exerts a statistically significant negative impact.

One way to see it is that when people know they've got help from AI in a task, they might not worry as much about how well they do on their own. Instead, they focus more on joining in competitive situations rather than aiming for personal success. This could be because they think the AI support makes up for any weaknesses they might have, making them more willing to take part in a competition without stressing too much about how good they are at the task (Gmyrek et al. 2023). People might trust that the AI help will make them better overall, giving them confidence that working with AI will cover any individual flaws (Dell'Acqua et al. 2023; Braganza et al. 2021 and Chong et al. 2022).

Table 4. Is there a gender gap when ChatGPT is utilized or selected? Probit estimates (marginal effects)

	[1]	[2]	[3]	[4]	[5]
Female	0.022 (0.034)	0.033 (0.034)	0.033 (0.035)	0.029 (0.036)	0.031 (0.036)
Tournament performance		-0.013*** (0.003)	-0.013*** (0.003)	-0.011*** (0.003)	-0.011*** (0.004)
Tournament- piece rate change in performance		0.012** (0.006)	0.011** (0.006)	0.013** (0.006)	0.012** (0.006)
Personality Traits			✓	✓	✓
Demographic and social economic variables				✓	✓
Psychological variables					✓
Pseudo R ²	0.010	0.015	0.018	0.060	0.063

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: N = 641. Dependent variable: Task-3 choice of tournament compensation scheme (1: tournament, 0: piece-rate). Tournament performance refers to Task-2 performance and tournament- piece rate change in performance refers to the change in performance between Task-2 and Task-1. The specifications control the experimental group and incentives (extrinsic/intrinsic).

*** p<0.01, ** p<0.05, *p<0.10.

Now we draw our attention to try to decompose these behavioral trends. Table 5 displays the impact of using AI and selecting AI (our treatment) on the likelihood of entering a tournament. We also consider interaction effects to better understand how these impacts differ based on gender. In Column (1) and Column (6), we examine the effects of ChatGPT utilization and ChatGPT selection, respectively. Interestingly, when individuals use ChatGPT without it being provided by their employer, there is a notable increase of 23.7 percentage points in the probability of entering a tournament compared to those who do not use it. In this scenario, the gender gap weakens, by decreasing to 6.1 p.p., and this effect is statistically significant at a 10% level. For those who select ChatGPT when it is offered as a tool, there is a higher probability of entering a tournament by 20.4 p.p. compared to those who do not opt for the treatment. What's intriguing is that, in this case, the gender gap completely disappears. To explain this finding, we examined interaction effects.

Columns (3)-(5) and (8)-(10) present the results while accounting for various factors such as performance variables, personality traits, demographics, socio-economic factors, and psychological variables. We provide specifications without controls to evaluate the overall impact of gender in the sample, recognizing that it may not be balanced across other features. However, our findings remain robust even after including demographic controls. The introduction of interaction effects alters the interpretation of coefficients. Regarding ChatGPT utilization, we observe that males using AI have a 18.5 p.p. higher likelihood of entering a tournament compared to males who don't use it. Similarly, females using AI have a 14.2 p.p. higher probability of entering a tournament compared to females who do not, although the effect is statistically weak (p-value < 0.10).

In the case of ChatGPT selection, it is noteworthy that while males show similar behavioral patterns to ChatGPT utilization, females opting for ChatGPT offered by the employer exhibit a 20.2 p.p. higher

probability of entering a tournament compared to females who do not select this option. This effect remains robust across all specifications (Column 7). This finding aligns with Hypothesis 2, suggesting that females place a higher value on seizing rewards offered by employers compared to males. This potential difference in perceived value could contribute to a greater willingness among females to engage with and leverage ChatGPT compared to their male counterparts.

Table 5. Decomposing gender gap in tournament entry when AI is utilized or selected. Probit estimates (marginal effects)

	ChatGPT utilization					ChatGPT Supply selection				
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Female	-0.061* (0.034)	-0.142*** (0.054)	-0.150*** (0.054)	-0.144*** (0.055)	-0.137** (0.057)	-0.023 (0.034)	-0.139*** (0.053)	-0.145*** (0.053)	-0.139*** (0.053)	-0.143*** (0.054)
AI utilization	0.237*** (0.041)	0.185*** (0.052)	0.197*** (0.052)	0.202*** (0.054)	0.215*** (0.057)					
Female \times utilization		0.142* (0.078)	0.145* (0.079)	0.145* (0.080)	0.131* (0.082)					
AI selection						0.204*** (0.041)	0.127*** (0.050)	0.123** (0.050)	0.121** (0.051)	0.130** (0.052)
Female \times selection							0.202*** (0.077)	0.227*** (0.077)	0.219*** (0.078)	0.220*** (0.080)
Performance variables			✓	✓	✓			✓	✓	✓
Personality Traits				✓	✓				✓	✓
Demographic and social economic variables					✓					✓
Psychological variables					✓					✓
Pseudo R ²	0.054	0.058	0.088	0.093	0.133	0.042	0.049	0.056	0.066	0.077
Observations			777					901		

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: Dependent variable: Task-3 choice of tournament compensation scheme (1: tournament, 0: piece-rate). The specifications control the experimental group and incentives (extrinsic/intrinsic). *** p<0.01, ** p<0.05, *p<0.10.

Overconfidence and Tournament Entry Choice

To elicit participants' beliefs on their relative tournament performance we asked them at the end of the experiment to guess how their performance in Task-2 ranked relative to the other members of their group. Participants earned \$0.50 if their estimation aligned with their actual ranking, and in cases of a tie, compensation was provided for any guesses considered accurate.

Relative to their actual rank, both men and women are overconfident. A Fisher's exact test of independence between the distribution of guessed rank and actual rank yields $p\text{-value} = 0.000$ for both men and women. However, men are more overconfident about their relative performance than women. While 35 percent of the men think they are best in their group of four, only 15 percent of the women hold this belief. The guesses of women and men differ significantly from one another, a Fisher's exact test of independence of the distributions for men and women delivers $p\text{-value} = 0.031$.

An ordered probit analysis of guessed rank, considering the influence of a female dummy variable and performance metrics (performance on Task-3 and change from Task-2 to Task-3), reveals an expected pattern. Even when controlling for performance, females, in general, express significantly lower confidence in their relative ranking compared to males (left panel) (Niederle & Vesterlund, 2007). Notably, the introduction of ChatGPT plays a crucial role in shaping this behavioral pattern of self-confidence. However, this influence is distinct based on the circumstances. In cases where females actively choose the employer-offered AI tool, a discernible impact on self-confidence is observed and the confidence difference by gender disappears (right panel). Surprisingly, when females independently opt for ChatGPT, this choice seems to have no notable effect on their self-confidence with females still being less confident than males. This is evident in their continued tendency to underrate their guessed ranking in the tournament compensation scheme, as depicted in Figure 6.

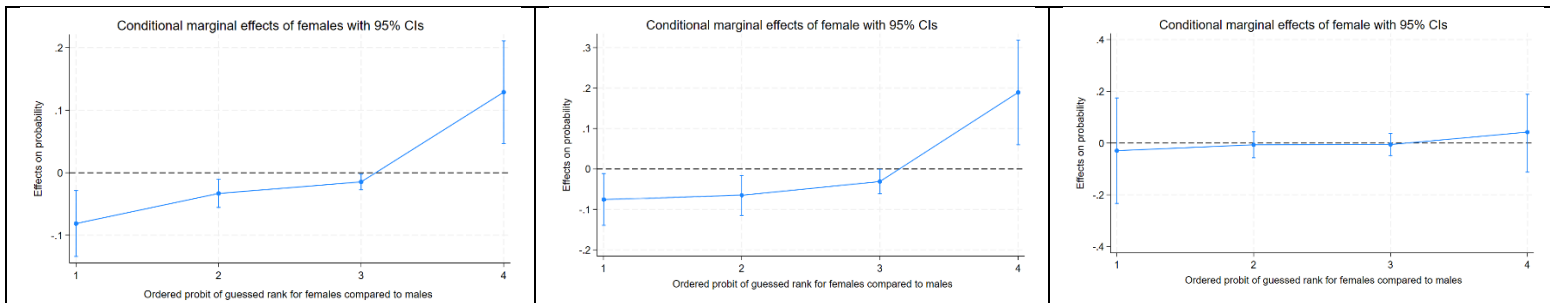


Figure 6. Marginal Effects of Guessed Rank for Females Compared to Males (1: First Place, 2: Second Place, 3: Third Place, 4: Fourth Place).

Left Panel: Overall sample results depicting the marginal effects of guessed rank.

Middle Panel: Marginal effects for participants who independently chose to utilize ChatGPT.

Right Panel: Results for individuals who opted for the ChatGPT treatment.

However, does the heightened overconfidence observed in females who opt for ChatGPT influence their competitive behavior, leading them to participate more frequently in tournaments? To examine this hypothesis, we introduced a binary variable named 'overconfidence', assigned a value of 1 for individuals predicting their ranking on 1st and 2nd place, and 0 otherwise. Subsequently, we incorporated this variable, along with a triple interaction term into our analysis to investigate the combined effects of females selecting ChatGPT and exhibiting overconfidence on the likelihood of entering the tournament. A series of probit regressions were conducted, and the results are presented in Table 6. Column (1) shows that participants, irrespective of their actual performance, exhibit a

greater likelihood of entering the tournament if they harbor higher confidence regarding their relative tournament performance in our comprehensive sample. However, by taking into consideration Figure 6 which reveals an absence of discernible effects of AI utilization on ranking guesses, in Columns (2)-(5), we exclude these participants, and we focus our analysis on individuals opting for ChatGPT and explore the interplay between AI selection and overconfidence on tournament entry, delineated by gender. Navigating through the specifications, we introduce a triple interaction term: *Female X AI Selection X Overconfidence*. Column (2) shows that – females embracing our AI supply treatment through ChatGPT, coupled with elevated levels of overconfidence, exhibit a 43.1 p.p. increase in the probability of tournament entry compared to their male counterparts. This convergence of female gender, AI selection, and overconfidence emerges as a pivotal factor influencing the dynamics of tournament participation.

Interestingly, this could be the reason why females who lean towards our ChatGPT treatment contribute to the narrowing of the gender gap in competitiveness. The finding suggests that when females choose ChatGPT provided by their employer, they also tend to feel more confident in their abilities. This boost in confidence encourages them to actively participate in competitive environments. This observation is in line with recent psychological studies that investigate why people utilize AI. For example, according to Fast & Schroeder (2020), using generative AI tools like ChatGPT can make individuals feel more empowered in their skills. This shift in mindset influences how they make decisions, as pointed out by De Freitas et al., (2023), who highlight the cognitive effects of interacting with such AI tools. This idea aligns with broader research indicating that when people interact with non-human technology, they feel less judged and are less hesitant to deviate from social norms, compared to interactions with real humans (Landers, 2019). In simpler terms, individuals seem to worry less about fitting into societal roles, like appearing competent, when interacting with AI. This creates an interesting connection between using AI, feeling overconfident, and actively participating in competitive settings, especially among females (Holthöwer & van Doorn, 2023).

Table 6. Overconfidence and Tournament entry (Probit marginal effects with third-degree interaction term)

	[1]	[2]	[3]	[4]	[5]
Female	-0.016 (0.031)	0.004 (0.004)	0.003 (0.004)	-0.011 (0.004)	-0.011 (0.004)
Overconfidence	0.676*** (0.023)	0.645*** (0.048)	0.646*** (0.048)	0.645*** (0.046)	0.655*** (0.046)
AI selection		0.178*** (0.049)	0.178*** (0.050)	0.200*** (0.051)	0.195*** (0.052)
AI selection X Overconfidence		0.745*** (0.019)	0.745*** (0.019)	0.755*** (0.019)	0.755*** (0.019)
Female X AI selection X Overconfidence		0.434*** (0.019)	0.431*** (0.019)	0.433*** (0.020)	0.431*** (0.020)
Performance variables	✓	✓	✓	✓	✓
Personality Traits			✓	✓	✓
Demographic and social economic variables				✓	✓

Psychological variables					✓
Pseudo R ²	0.190	0.241	0.252	0.267	0.269
Observations	1216			901	

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: Dependent variable: Task-3 choice of tournament compensation scheme (1: tournament, 0: piece-rate). The specifications control the experimental group and incentives (extrinsic/intrinsic).

*** p<0.01, ** p<0.05, *p<0.10.

Risk Aversion on Past-Performance (Task-4) and Tournament Entry Choice

We use Task-4 to examine whether a females and males tournament entry behavior are difference when the tournament choice does not require a subsequent competitive performance. Participants in Task 4 had the opportunity to select one of two compensation schemes for their past piece-rate performance on Task-1, either the piece rate or the tournament. If the tournament is chosen, the piece-rate performance is submitted to a competition against the piece-rate performances of the other participants in the group (independent of their choice of compensation scheme). A tournament is won if an individual's performance exceeds that of the other three players. In our analysis this variable is coded as 1 if the participant's selection is the past piece-rate performance and 0 otherwise. Hence Task-4 can serve as a proxy of risk aversion, with individuals making the choice of part performance on piece rate being considered as more risk averted (Niederle & Vesterlund, 2007).

Firstly, in general, we do not observe a statistically significant difference in Task-4 choice by gender (paired t-test, p-value = 0.892). However, we do find that individuals who either utilize or select AI have a 30 p.p. and 19 p.p. lower probability, respectively, of choosing this compensation scheme option. This is conditional on their performance variables (p-value = 0.00, probit marginal effects).

To assess the joint effect of this compensation scheme choice, AI and gender we again introduce a triple interaction term: *Female X AI Utilization/Selection X Past Piece-Rate performance choice*. The results are presented in Table 7, with Panel A including participants who independently utilize ChatGPT, and Panel B consisting of individuals who select the ChatGPT treatment. In Column (1), we examine the total sample and find that participants who opt for their past piece-rate performance are generally not expected to choose tournament entry. However, these effects are not statistically significant. Columns (2)-(5) focus on each panel separately. Notably, in both cases, we do not observe statistically significant effects for females utilizing or selecting ChatGPT while also being risk-averse (choosing past performance on Task-1 as their compensation scheme). Interestingly, it appears that ChatGPT reverses the effect of risk aversion observed in Column (1).

Table 7. Past Piece-Rate performance choice and Tournament entry (Probit marginal effects with third-degree interaction term)

	[1]	[2]	[3]	[4]	[5]
	Panel A: ChatGPT utilization				
Female	-0.033 (0.028)	-0.054 (0.036)	-0.049 (0.036)	-0.056 (0.037)	-0.052 (0.037)
Past Piece-Rate performance choice	-0.048 (0.032)	-0.126** (0.053)	-0.124** (0.053)	-0.121** (0.056)	-0.108* (0.058)
AI utilization		0.191*** (0.048)	0.197*** (0.050)	0.230*** (0.052)	0.217*** (0.053)
AI utilization X Past Piece-Rate performance choice		0.289*** (0.104)	0.268** (0.108)	0.220** (0.113)	0.198* (0.114)

Female \times AI utilization \times Past Piece-Rate performance choice		-0.177* (0.092)	-0.170* (0.095)	-0.161* (0.095)	-0.154 (0.097)
Pseudo R ²	0.013	0.093	0.099	0.131	0.136
Observations	1216		777		
Panel B: ChatGPT Supply selection					
Female	-0.033 (0.028)	-0.012 (0.036)	-0.011 (0.036)	-0.017 (0.037)	-0.014 (0.036)
Past Piece-Rate performance choice	-0.048 (0.032)	-0.121** (0.051)	-0.121** (0.052)	-0.122** (0.053)	-0.115** (0.055)
AI selection		0.157*** (0.047)	0.154*** (0.048)	0.170*** (0.048)	0.168*** (0.048)
AI selection \times Past Piece-Rate performance choice		0.283*** (0.102)	0.280*** (0.103)	0.262** (0.107)	0.259** (0.108)
Female \times AI selection \times Past Piece-Rate performance choice		-0.077 (0.098)	-0.076 (0.099)	-0.081 (0.100)	-0.083 (0.099)
Pseudo R ²	0.013	0.055	0.058	0.074	0.078
Observations	1216		901		
Performance variables	✓	✓	✓	✓	✓
Personality Traits			✓	✓	✓
Demographic and social economic variables				✓	✓
Psychological variables					✓

Source: Authors' calculations. Data drawn from the Amazon Mechanical Turk experiment.

Notes: Dependent variable: Task-3 choice of tournament compensation scheme (1: tournament, 0: piece-rate). The specifications control the experimental group and incentives (extrinsic/intrinsic).

*** p<0.01, ** p<0.05, *p<0.10.

V. DISCUSSION

This paper presents the findings of an experiment examining the impact of integrating an AI tool on individuals' inclination to participate in a competition. Employing an online task proven to be gender-neutral in measured performance, we observe that females are equally likely as men to enter the tournament when they select our ChatGPT treatment. Despite an 11-percentage-point gender gap observed for individuals not utilizing ChatGPT, this gap disappears when ChatGPT is introduced into the scenario.

A key finding is the observation of distinct behavioral patterns between females who independently choose to utilize ChatGPT and those who select ChatGPT when it is offered by the employer-experimenter. In the former case, the gender gap decreases, while in the latter, it completely vanishes. To explore this mechanism, we focus on confidence patterns, given previous studies indicating that females tend to appear less confident than males (Niederle & Vesterlund, 2007; Charness et al., 2018; Jakobsson et al., 2013). Intriguingly, we discover that only in the case of females opting for ChatGPT in the treatment group do they exhibit overconfidence, resulting in an increased trend in tournament entry rates. Further analysis on gender differences in risk aversion does not provide strong evidence for systematic variations in the link between ChatGPT and risk aversion by gender. Another notable observation is that ChatGPT appears to enhance the performance of women in the task, with no significant effect on men. While increased performance in women does not necessarily guarantee a rise in competitiveness intention (Charness et al., 2022), in our case, it suggests that women may

strategically respond optimally. This is evidenced by their choice of ChatGPT, increased performance, and apparent overconfidence, all influencing their decisions regarding competitiveness.

In attempting to explain why these behavioral patterns are mainly found in cases where females select ChatGPT when it is offered, our survey results indicate that female beliefs toward AI-tool use in the workplace center around signaling misbehavior and the need for competence. Additionally, personality traits and the honesty-humility factors, show evidence that ChatGPT utilization is perceived as a signal for misbehavior and cheating, with pronounced negative effects on conscientiousness and fairness, when not officially offered by the employer. Notably, these effects dissipate when ChatGPT is formally provided by the employer. Building on prior research, we posit that in our context, females who select ChatGPT may employ it as an educational aid to enhance their preparation for upcoming job challenges (Charness et al., 2022; Zhang et al., 2019). Additionally, they might place a greater emphasis on capitalizing on rewards provided by employers compared to males within the workplace setting (Avery et al., 2023).

Considering our findings that highlight the significant role of AI in workplaces and its potential to reduce the gender gap in competitiveness, several policy implications come to the forefront. For example, policymakers should support the inclusive adoption of AI tools across all professional spheres, prioritizing accessibility, and training opportunities for employees, irrespective of gender. Tailored training programs addressing gender-specific barriers can further empower individuals to be equipped with the necessary skills allowing them not only to make use of AI technologies, but to contribute to their gender-specific development¹⁶.

Future research should explore deeper the intricate connection between the utilization of generative AI-tools and behavioral outcomes, exploring various dimensions of human-computer interaction. For instance, investigating the role of AI in mixed-gender teams and its impact on collaborative competitiveness is crucial. Furthermore, understanding how AI usage can contribute to the evolution of cognitive strategies, including strategic thinking, problem-solving, and decision-making skills, with the potential to mitigate gender inequalities in the workplace warrants exploration. Ethical considerations surrounding AI use, particularly in decision-making, present another avenue for investigation. Lastly, examining how minorities can ensure workplace resilience through the integration of AI adds a valuable dimension to the research agenda.

VI. CONCLUSION

It is crucial to acknowledge the manifold impacts that emerging technologies can exert on labor markets. Our study explores the ramifications of adopting generative AI on the gender gap. Notably, our findings reveal that when women incorporate ChatGPT as a tool in their work, their competitiveness intentions witness a notable uptick. This pivotal discovery underscores the potential of generative AI-tools to facilitate the integration of minority groups into the workforce on equal terms. This research sheds light on the prospect of dismantling psychological barriers that women often encounter when venturing into traditionally male-dominated professions, especially with the introduction of AI-tools. The influence of generative AI-tools is not limited to shaping women's decision to compete but extends to various environments marked by existing barriers in the labor market. For instance, the implementation of AI-tools could aid employers in assessing and enhancing

¹⁶ There is growing evidence nowadays of gender bias embedded in AI tools, often favoring a male-centric perspective (Manasi et al., 2022).

the career advancement prospects of female employees. However, further research is imperative to comprehensively gauge the extent of these potential transformations.

Authorship contribution statement

Evangelos Mourelatos: Conceptualization, Methodology, Data analysis, Experimental Design, Software Writing – original draft, Writing – review & editing, Project administration. **Panagiotis Zervas:** Investigation, Data curation, Experimental Design, Writing – review & editing. **Dimitris Lagios:** Data acquisition, Programming. **Giannis Tzimas:** Project administration, Writing – review & editing.

References

- Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial intelligence and jobs: evidence from online vacancies. *Journal of Labor Economics*, 40(S1), S293-S340.
- Acemoglu, D. (2021). Harms of AI (No. w29247). National Bureau of Economic Research.
- Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Artificial intelligence: the ambiguous labor market impact of automating prediction. *Journal of Economic Perspectives*, 33(2), 31-50.
- Andersen, S., Ertac, S., Gneezy, U., List, J. A., & Maximiano, S. (2013). Gender, competitiveness, and socialization at a young age: Evidence from a matrilineal and a patriarchal society. *Review of Economics and Statistics*, 95(4), 1438-1443.
- Anderson, N., Belavy, D., Perle, S., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation. *BMJ Open Sport — Exercise Medicine*.
- Apicella, C. L., & Dreber, A. (2015). Sex differences in competitiveness: Hunter-gatherer women and girls compete less in gender-neutral and male-centric tasks. *Adaptive Human Behavior and Physiology*, 1, 247-269.
- Apicella, C. L., Demiral, E. E., & Mollerstrom, J. (2017). No gender difference in willingness to compete when competing against self. *American Economic Review*, 107(5), 136-140.
- Apostolou, M., & Panayiotou, R. (2019). The reasons that prevent people from cheating on their partners: An evolutionary account of the propensity not to cheat. *Personality and Individual Differences*, 146, 34-40.
- Arlow, P. (1991). Personal characteristics in college students' evaluations of business ethics and corporate social responsibility. *Journal of Business Ethics*, 10, 63-69.
- Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, 18(2), 139-152.
- Avery, M., Leibbrandt, A., & Vecchi, J. (2023). Does Artificial Intelligence Help or Hurt Gender Diversity? Evidence from Two Field Experiments on Recruitment in Tech (No. 2023-09). Monash University, Department of Economics. Working Paper
- Autor, D. (2022). The labor market impacts of technological change: From unbridled enthusiasm to qualified optimism to vast uncertainty (No. w30074). National Bureau of Economic Research.
- Awad, E., Balafoutas, L., Chen, L., Ip, E., & Vecchi, J. (2023). Artificial Intelligence and Debiasing in Hiring: Impact on Applicant Quality and Gender Diversity. Available at SSRN.
- Babina, T., Fedyk, A., He, A. X., & Hodson, J. (2022). Firm investments in artificial intelligence technologies and changes in workforce composition. Available at SSRN 4060233.
- Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics*, 151, 103745.
- Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), 1-17.

- Bao, Z, D Huang, C Lin (2024), "Can Artificial Intelligence Improve Gender Equality? Evidence from a Natural Experiment" *Management Science*, forthcoming.
- Barber, B. M., & Odean, T. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *The Quarterly Journal of Economics*, 116(1), 261-292.
- Beyer, S. (1990). Gender differences in the accuracy of self-evaluations of performance. *Journal of Personality and Social Psychology*, 59(5), 960.
- Beyer, S., & Bowden, E. M. (1997). Gender differences in self-perceptions: Convergent evidence from three measures of accuracy and bias. *Personality and Social Psychology Bulletin*, 23(2), 157-172.
- Bin-Nashwan, S. A., Sadallah, M., & Bouteraa, M. (2023). Use of ChatGPT in academia: Academic integrity hangs in the balance. *Technology in Society*, 75, 102370.
- Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature*, 55(3), 789-865.
- Booth, A., & Nolen, P. (2012). Choosing to compete: How different are girls and boys?. *Journal of Economic Behavior & Organization*, 81(2), 542-555.
- Bowles, H. R., & Gelfand, M. (2010). Status and the evaluation of workplace deviance. *Psychological Science*, 21(1), 49-54.
- Boyce, W., Torsheim, T., Currie, C., & Zambon, A. (2006). The family affluence scale as a measure of national wealth: validation of an adolescent self-report measure. *Social indicators research*, 78, 473-487.
- Braganza, A., Chen, W., Canhoto, A., & Sap, S. (2021). Productive employment and decent work: The impact of AI adoption on psychological contracts, job engagement and employee trust. *Journal of Business Research*, 131, 485-494.
- Braguinsky, S., Ohyama, A., Okazaki, T., & Syverson, C. (2021). Product innovation, product diversification, and firm growth: Evidence from Japan's early industrialization. *American Economic Review*, 111(12), 3795-3826.
- Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Burbano, V. C., Folke, O., Meier, S., & Rickne, J. (2023). The Gender Gap in Meaningful Work. *Management Science*.
- Buser, T., van den Assem, M. J., & van Dolder, D. (2023). Gender and willingness to compete for high stakes. *Journal of Economic Behavior & Organization*, 206, 350-370.
- Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, competitiveness, and career choices. *The Quarterly Journal of Economics*, 129(3), 1409-1447.
- Carlsson, F., Lampi, E., Martinsson, P., & Yang, X. (2020). Replication: Do women shy away from competition? Experimental evidence from China. *Journal of Economic Psychology*, 81, 102312.

- Chadi, A., & Homolka, K. (2022). Little lies and blind eyes—Experimental evidence on cheating and task performance in work groups. *Journal of Economic Behavior & Organization*, 199, 122-159.
- Charness, G., Rustichini, A., & Van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics*, 21, 72-98.
- Charness, G., Jabarian, B., & List, J. A. (2023). Generation next: Experimentation with ai (No. w31679). National Bureau of Economic Research.
- Charness, G., Dao, L., & Shurchkov, O. (2022). Competing now and then: The effects of delay on competitiveness across gender. *Journal of Economic Behavior & Organization*, 198, 612-630.
- Chen MK, Rossi PE, Chevalier JA, Oehlsen E (2019). The value of flexible work: Evidence from uber drivers. *Journal of Political Economy* 127(6):2735–2794.
- Chen, M. K., & Chevalier, J. A. (2012). Are women overinvesting in education? Evidence from the medical profession. *Journal of Human Capital*, 6(2), 124-149.
- Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior*, 127, 107018.
- Claridge, G., & Davis, C. (2001). What's the use of neuroticism?. *Personality and Individual Differences*, 31(3), 383-400.
- Compte, O., & Postlewaite, A. (2004). Confidence-enhanced performance. *American Economic Review*, 94(5), 1536-1557.
- Costa Jr, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: robust and surprising findings. *Journal of Personality and Social Psychology*, 81(2), 322.
- Costa, P. T., & McCrae, R. R. (1999). A five-factor theory of personality. *The five-factor model of personality: Theoretical Perspectives*, 2, 51-87.
- Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2), 448-474.
- Cubel, M., Nuevo-Chiquero, A., Sanchez-Pages, S., & Vidal-Fernandez, M. (2016). Do personality traits affect productivity? Evidence from the laboratory. *The Economic Journal*, 126(592), 654-681.
- Datta Gupta, N., Poulsen, A., & Villeval, M. C. (2013). Gender matching and competitiveness: Experimental evidence. *Economic Inquiry*, 51(1), 816-835.
- De Freitas, J., Agarwal, S., Schmitt, B., & Haslam, N. (2023). Psychological factors underlying attitudes toward AI tools. *Nature Human Behaviour*, 7(11), 1845-1854.
- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... & Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).
- Di Gangi, P. M., McAllister, C. P., Howard, J. L., Thatcher, J. B., & Ferris, G. R. (2022). Can you see opportunity knocking? An examination of technology-based political skill on opportunity recognition in online communities for MTurk workers. *Internet Research*, 32(4), 1041-1075.

- El-Sayed Abd-Elaal, Sithara H.P.W. Gamage & Julie E. Mills (2022). Assisting academics to identify computer generated writing, *European Journal of Engineering Education*, 47:5, 725-745.
- Fast, N. J., & Schroeder, J. (2020). Power and decision making: new directions for research in the age of artificial intelligence. *Current opinion in psychology*, 33, 172-176.
- Felten, E. W., Raj, M., & Seamans, R. (2018). A method to link advances in artificial intelligence to occupational abilities. In *AEA Papers and Proceedings* (Vol. 108, pp. 54-57). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.
- Felten E, Raj M, Seamans R (2021) Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal* 42(12):2195–2217.
- Felten, E., Raj, M., & Seamans, R. (2023). How will Language Modelers like ChatGPT Affect Occupations and Industries?. *arXiv preprint arXiv:2303.01157*.
- Flory, J. A., Leibbrandt, A., & List, J. A. (2015). Do competitive workplaces deter female workers? A large-scale natural field experiment on job entry decisions. *The Review of Economic Studies*, 82(1), 122-155.
- Gallego A, Kurer T (2022) Automation, digitalization, and artificial intelligence in the workplace: implications for political behavior. *Annual Review of Political Science* 25:463–484.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.
- Gill, D., & Prowse, V. (2014). Gender differences and dynamics in competition: The role of luck. *Quantitative Economics*, 5(2), 351-376.
- Giluk, T. L., & Postlethwaite, B. E. (2015). Big Five personality and academic dishonesty: A meta-analytic review. *Personality and Individual Differences*, 72, 59-67.
- Gmyrek, P., Berg, J., & Bescond, D. (2023). Generative AI and Jobs: A global analysis of potential effects on job quantity and quality. *ILO Working Paper*, 96.
- Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, 118(3), 1049-1074.
- Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., & Brauner, J. (2024). Thousands of AI Authors on the Future of AI. *arXiv preprint arXiv:2401.02843*.
- Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In *Handbook of research methods in consumer psychology* (pp. 319-337). Routledge.
- Hilbig, B. E., Heydasch, T., & Zettler, I. (2014). To boast or not to boast: Testing the humility aspect of the Honesty–Humility factor. *Personality and Individual Differences*, 69, 12-16.

- Hillebrandt, A., & Barclay, L. J. (2020). How cheating undermines the perceived value of justice in the workplace: The mediating effect of shame. *Journal of Applied Psychology*, 105(10), 1164.
- Holthöwer, J., & van Doorn, J. (2023). Robots do not judge: service robots can alleviate embarrassment in service encounters. *Journal of the Academy of Marketing Science*, 51(4), 767-784.
- Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? (No. w31122). National Bureau of Economic Research.
- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14, 399-425.
- Huffman, A. H., Whetten, J., & Huffman, W. H. (2013). Using technology in higher education: The influence of gender roles on technology self-efficacy. *Computers in Human Behavior*, 29(4), 1779-1786.
- Hui, X., Reshef, O., & Zhou, L. (2023). The short-term effects of generative artificial intelligence on employment: Evidence from an online labor market. Available at SSRN 4527336.
- Instone, D., Major, B., & Bunker, B. B. (1983). Gender, self-confidence, and social influence strategies: An organizational simulation. *Journal of Personality and Social Psychology*, 44(2), 322.
- Ipeirotis, P. G. (2010). Demographics of mechanical turk.
- Jakobsson, N., Levin, M., & Kotsadam, A. (2013). Gender and overconfidence: effects of context, gendered stereotypes, and peer group. *Advances in Applied Sociology*, 3(02), 137.
- John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, Measurement, and Theoretical Perspectives.
- Johnson, M. K., Rowatt, W. C., & Petrini, L. (2011). A new trait on the market: Honesty–Humility as a unique predictor of job performance ratings. *Personality and Individual differences*, 50(6), 857-862.
- Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.
- Kazai, G., Kamps, J., & Milic-Frayling, N. (2011). Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1941-1944).
- Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2583-2586).
- Klinowski, D. (2019). Selection into self-improvement and competition pay: Gender, stereotypes, and earnings volatility. *Journal of Economic Behavior & Organization*, 158, 128-146.
- Kim, A., Muhn, M., & Nikolaev, V. (2023). From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI. arXiv preprint arXiv:2310.17721.
- Kokkodis M, Ransbotham S (2023) Learning to successfully hire in online labor markets. *Management Science* 69(3):1597–1614.

- Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4), 1281-1317.
- Koszegi, Botond, "Ego Utility, Overconfidence, and Task Choice," *Journal of the European Economic Association*, 2006, 4 (4), 673–707.
- Landers, R. N. (Ed.). (2019). *The Cambridge handbook of technology and employee behavior*. Cambridge University Press.
- Leinonen, J., Longi, K., Klami, A., Ahadi, A., & Vihavainen, A. (2016). Typing Patterns and Authentication in Practical Programming Exams. *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*.
- Lenney, E. (1977). Women's self-confidence in achievement settings. *Psychological Bulletin*, 84(1), 1.
- List, J. A., & Momeni, F. (2020). Leveraging upfront payments to curb employee misbehavior: Evidence from a natural field experiment. *European Economic Review*, 130, 103601.
- Liu, J., Xu, X., Li, Y., & Tan, Y. (2023). "Generate" the Future of Work through AI: Empirical Evidence from Online Labor Markets. *arXiv preprint arXiv:2308.05201*.
- Luo, X., Qin, M. S., Fang, Z., & Qu, Z. (2021). Artificial intelligence coaches for sales agents: Caveats and solutions. *Journal of Marketing*, 85(2), 14-32.
- Lysyakov, M., & Viswanathan, S. (2023). Threatened by AI: Analyzing Users' Responses to the Introduction of AI in a Crowd-sourcing Platform. *Information Systems Research*, 34(3), 1191-1210.
- Manasi, A., Panchanadeswaran, S., Sours, E., & Lee, S. J. (2022). Mirroring the bias: gender and artificial intelligence. *Gender, Technology and Development*, 26(3), 295-305.
- Mander, A. (2019). *SAMPSI_REG: Stata module to calculate the sample size/power for linear regression*.
- Marciano, L., Camerini, A. L., & Schulz, P. J. (2020). Neuroticism in the digital age: A meta-analysis. *Computers in Human Behavior Reports*, 2, 100026.
- Markowsky, E., & Beblo, M. (2022). When do we observe a gender gap in competition entry? A meta-analysis of experimental literature. *Journal of Economic Behavior & Organization*, 198, 139-163.
- McCrae, R. R. (1987). Creativity, divergent thinking, and openness to experience. *Journal of Personality and Social Psychology*, 52(6), 1258.
- Miesing, P., & Preble, J. F. (1985). A comparison of five business philosophies. *Journal of business ethics*, 4, 465-476.
- Moore, D. A., & Schatz, D. (2017). The three faces of overconfidence. *Social and Personality Psychology Compass*, 11(8), e12331.
- Mourelatos, E., Giannakopoulos, N., & Tzagarakis, M. (2022). Personality traits and performance in online labour markets. *Behaviour & Information Technology*, 41(3), 468-484.

- Mourelatos, E. (2023). Does Mood affect Sexual and Gender Discrimination in Hiring Choices? Evidence from Online Experiments. *Journal of Behavioral and Experimental Economics*, 106, 102069.
- Mourelatos, E., Giannakopoulos, N., & Tzagarakis, M. (2023). Payment schemes in online labour markets. Does incentive and personality matter?. *Behaviour & Information Technology*, 1-22.
- Newson, R. (2004). Generalized power calculations for generalized linear models and more. *The Stata Journal*, 4(4), 379-401.
- Niederle, M., & Vesterlund, L. (2007). Do women shy away from competition? Do men compete too much?. *The Quarterly Journal of Economics*, 122(3), 1067-1101
- Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 2023, 381(6654), p.187-192.
- Organisation for Economic Co-operation and Development, (2019). *Artificial Intelligence in Society*. OECD Publishing, Paris.
- Parker, S. K., & Grote, G. (2022). Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Applied Psychology*, 71(4), 1171-1204.
- Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46, 1023-1031.
- Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching & Learning Practice*, 20(2).
- Qiao, D., Rui, H., & Xiong, Q. (2023). AI and Jobs: Has the Inflection Point Arrived? Evidence from an Online Labor Platform. *arXiv preprint arXiv:2312.04180*.
- Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, 27(2), 151-161.
- Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package*, 61(52), 18.
- Roth AE. (2015) *Who gets what—and why: the new economics of matchmaking and market design* (Houghton Mifflin Harcourt).
- Roth, J., Liu, X., & Metaxas, D. (2014). On Continuous User Authentication via Typing Behavior. *IEEE Transactions on Image Processing*, 23, 4611-4624.
- Sackett, P. R., & Wanek, J. E. (1996). New developments in the use of measures of honesty integrity, conscientiousness, dependability trustworthiness, and reliability for personnel selection. *Personnel Psychology*, 49(4), 787-829.
- Schneider, T. R. (2004). The role of neuroticism on psychological and physiological stress responses. *Journal of Experimental Social Psychology*, 40(6), 795-804.

- Segovia-Pérez, M., Castro Núñez, R. B., Santero Sánchez, R., & Laguna Sánchez, P. (2020). Being a woman in an ICT job: an analysis of the gender pay gap and discrimination in Spain. *New Technology, Work and Employment*, 35(1), 20-39.
- Shurchkov, O. (2012). Under pressure: gender differences in output quality and quantity under competition and time constraints. *Journal of the European Economic Association*, 10(5), 1189-1213.
- Sinning, M. (2017). Gender differences in costs and returns to higher education. *AND GENDER*, 227.
- Stoet, G., & Geary, D. C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science*, 29(4), 581-593.
- Taddy, M. (2018). The technological elements of artificial intelligence. In *The economics of artificial intelligence: An agenda* (pp. 61-87). University of Chicago Press.
- Tyson, T. (1992). Does believing that everyone else is less ethical have an impact on work behavior?. *Journal of Business Ethics*, 11, 707-717.
- van Inwegen, E., Munyikwa, Z. T., & Horton, J. J. (2023). Algorithmic writing assistance on jobseekers' resumes increases hires (No. w30886). National Bureau of Economic Research.
- van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness?. *Journal of the European Economic Association*, 20(4), 1595-1618.
- Venkatesh, V., Morris, M. G., Sykes, T. A., & Ackerman, P. L. (2004). Individual reactions to new technologies in the workplace: The role of gender as a psychological construct. *Journal of Applied Social Psychology*, 34(3), 445-467.
- Wang Y, Kordi Y, Mishra S, Liu A, Smith NA, Khashabi D, Hajishirzi H (2022) Self-instruct: Aligning language model with self generated instructions. arXiv preprint arXiv:2212.10560.
- Webb, M. (2019). The impact of artificial intelligence on the labor market. Available at SSRN 3482150.
- Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The menstrual cycle and performance feedback alter gender differences in competitive choices. *Journal of Labor Economics*, 32(1), 161-198.
- Yilmaz, E. D., Naumovska, I., & Aggarwal, V. A. (2023). AI-Driven Labor Substitution: Evidence from Google Translate and ChatGPT. Available at SSRN.
- Young, E., Wajcman, J., & Sprejer, L. (2023). Mind the gender gap: Inequalities in the emergent professions of artificial intelligence (AI) and data science. *New Technology, Work and Employment*, 38(3), 391-414.
- Young, E., Wajcman, J., & Sprejer, L. (2021). Where are the women? Mapping the gender job gap in AI. Working paper.
- Zhang, Y., Xu, Z., & Palma, M. A. (2019). Conveniently dependent or naively overconfident? An experimental study on the reaction to external help. *Plos one*, 14(5), e0216617.

Zheng, H., Li, D., & Hou, W. (2011). Task design, motivation, and participation in crowdsourcing contests. *International Journal of Electronic Commerce*, 15(4), 57-88.