ISSN (Print) 0473-453X
Discussion Paper No. 1233
ISSN (Online) 2435-0982

DO PEOPLE RELY ON ChatGPT MORE THAN THEIR PEERS TO DETECT DEEPFAKE NEWS?

Yuhao Fu
Nobuyuki Hanaki

Revised December 2024
March 2024

The Institute of Social and Economic Research
Osaka University
6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan

Do people rely on ChatGPT more than their peers to detect deepfake news?*

Yuhao Fu† Nobuyuki Hanaki‡

December 5, 2024

Abstract

This experimental study investigates whether people rely more on ChatGPT (GPT-4) than on their human peers when detecting AI-generated fake news (deepfake news). In multiple rounds of deepfake detection tasks conducted in a laboratory setting, student participants exhibited a greater reliance on ChatGPT compared to their peers. We explored this over-reliance on AI from two perspectives: the weight of advice (WOA) and the decomposition of reliance (DOR) into two stages. Our analysis indicates that reliance on external advice is primarily influenced by the source and quality of the advice, as well as the subjects' prior beliefs, knowledge, and experience, while the type of news and time spent on tasks have no effect. Additionally, our study indicates a potential sequential mechanism of advice utilization, wherein the advice source affects reliance in both stages—activation and integration—whereas the quality of the advice, along with knowledge and experience, influences only the second stage. Our findings suggest that relying on AI to detect AI may not be detrimental and could, in fact, contribute to a deeper understanding of human-AI interaction and support advancements in AI development during the Generative Artificial Intelligence (GAI) era.

Keywords: ChatGPT, AI reliance, deepfake detection, weight of advice (WOA), decomposition of reliance (DOR)

JEL: C90; D83; D90; D91

*This research has benefited from the financial support of (a) the Joint Usage/Research Center, the Institute of Social and Economic Research (ISER), and Osaka University, and (b) Grants-in-aid for Scientific Research Nos. 20H05631 and 23H00055 from the Japan Society for the Promotion of Science. The design of the experiment reported in this paper was approved by the IRB of ISER (#20231001) in October 2023, and the experiment is preregistered at aspredicted.org (#149838). We also want to thank Tiffany Tsz Kwan Tse and Taisuke Imai for their helpful comments and suggestions. Supports from Yuta Shimodaira, Zhelin Zhang, and Satsuki Yamada for conducting the experiment is also gratefully acknowledged.

†Graduate School of Economics, Osaka University. E-mail: u889037j@ecs.osaka-u.ac.jp

‡Corresponding author. Institute of Social and Economic Research, Osaka University, and University of Limassol. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

1 Introduction

GAI products, such as ChatGPT, have gained worldwide attention since 2022 (Zhang et al., 2023; Aydın and Karaarslan, 2023). By creating new content—from images to dialogues—simply based on user inputs, GAI enhances AI's integration into daily life. Consequently, concerns about people's over-reliance on AI have been increasing rapidly (LaGrandeur, 2021; Cao and Huang, 2022; Schemmer et al., 2023; Klingbeil et al., 2024). However, GAI also generates potential risks, particularly by contributing to the spread of AI-generated fake news (Fui-Hoon Nah et al., 2023; Walkowiak and MacDonald, 2023), which can also be called "deepfake news" (Whyte, 2020; Lee and Shin, 2022). This not only raises concerns about people's ability to detect deepfake content (Bray et al., 2023; Mai et al., 2023), but also poses additional challenges to deepfake detection technologies (Wang, 2024; Takale et al., 2024). Recognizing this issue, researchers have suggested leveraging GAI itself to combat these risks (Patil et al., 2024; Bhattacharjee and Liu, 2024). As a result, GAI has now become both a source of and a tool against deepfake content. This duality underscores the importance of understanding how individuals rely on AI when confronted with AI-generated misinformation.

With this context in mind, this study seeks to better characterize people's reliance on AI, and their perception of the deepfake contents generated by AI. We introduced a deepfake detection task in the laboratory, where participants were asked to assess the proportion of human-written content in synthetic news articles, which were compositions of human-written and AI-generated (GPT-2) text. To investigate the reliance, we applied a between-subject design, providing participants with assessments from either ChatGPT (GPT-4) or their human peers as advice to help them adjust their initial judgments. By using this comparative setting, we established human peers as a relative baseline for AI, thereby obtaining a concrete understanding of AI reliance. Our analysis mainly revealed that, participants relied more on ChatGPT then their human peers

in the special task of deepfake detection. This finding was demonstrated using two methods: the "weight of advice" (WOA) and the "decomposition of reliance" (DOR).

The first method, WOA, is a common metric in the psychology of advice utilization to measure the degree to which people take advice (Harvey and Fischer, 1997; Yaniv and Foster, 1997) in the form of a numerical estimate. Some studies directly quantify people's reliance on advice sources using this approach (Önkal et al., 2009; Castelo et al., 2019). In our analysis, we mainly compared the WOA between different groups where the advice source varied (human peers vs. ChatGPT) and also investigated the effects of advice quality, news types, subjects' prior beliefs and time spent on tasks. We observed that the WOA is significantly higher when the advice source is ChatGPT compared to human peers, indicating greater reliance on AI's advice. Additionally, the degree of reliance is also affected by advice quality, subjects' prior beliefs, knowledge and experience in programming.

In the analysis of DOR, we employed the two-stage model of Vodrahalli et al. (2022) along with the Heckman correction (Heckman, 1974; Heckman and RajBhandary, 1979) to decompose reliance into two stages: "activation" and "integration". In detail, during the first stage (activation), participants determine whether to accept the advice and update their initial judgments. Then, in the second stage (integration), if they choose to update, they decide the extent to which they will utilize the advice. Our results indicate that the Heckman correction confirms the feasibility of this decomposition. We found that "AI reliance" and the effect of prior beliefs are present in both stages. However, the advice's quality, along with participants' knowledge and experience in programming, did not affect the activation stage but did influence the integration stage. This finding illustrates a sequential mechanism in people's process of utilization. Specifically, they separate the source of the advice from its content. Only after being "activated" by the source and deciding to accept the advice do they consider the specific quality of the advice by integrating their own knowledge and experience.

In addition to examining reliance, we also discussed participants' performance in the deepfake detection task. This type of task is often used in studies of AI model development, especially as a part of human evaluation conducted online (Agarwal et al., 2019; Tolosana et al., 2020; Nguyen et al., 2022), but is rarely used in economic experiments. Since participants' payoffs depend on their performance in the task and our primary focus is on reliance, we conducted this experiment in the laboratory to prevent participants from accessing other advice sources. The results show that participants had an overall accuracy of about 70.8% when not receiving external advice (from human peers or ChatGPT). After receiving advice, participants who received ChatGPT's advice did not outperform those who received peers' advice. Although both the AI advisor and human advisor provided advice of similar quality, participants receiving ChatGPT's advice showed greater improvement compared to those receiving advice from human peers. This suggests that higher reliance on AI translates to improved outcomes, indicating that the reliance on AI for detecting (AI-generated) deepfake news may not be detrimental.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on deepfake detection and AI reliance. Section 3 presents the experimental design and hypotheses. Section 4 summarizes the analysis results of WOA, while Section 5 presents the results of DOR analysis. Section 6 provides a discussion of the findings, and Section 7 concludes the paper.

2 Literature Review

Human reliance on nonhuman systems has often been studied in the fields of psychology and economics. Among related studies, researchers frequently compare how humans take advice from AI versus humans using the judge-advisor paradigm (JAS) (Sniezek and Buckley, 1989). In this framework, participants are asked to complete

a forecasting task, provide an initial response, subsequently receive advice from external sources, and then offer a second response. This allows researchers to measure reliance on advice sources by assessing the extent to which participants update their initial responses based on the advice provided. Previous studies have employed this approach across various forecasting tasks, including historical dates (Gino, 2008), stock prices (Önkal et al., 2009), song popularity (Logg et al., 2019), object weight estimation (Yoon et al., 2021), and others.

However, this study introduces a deepfake detection task, which is rarely used to investigate human reliance on external advice sources. Therefore, in the following subsections, we will review background of deepfake detection to elucidate its importance for our study and subsequently examine previous work on AI reliance.

2.1 Background on Deepfake Detection

Deepfakes refer to synthetic media—primarily audio and video—that have been generated using deep learning techniques (Chadha et al., 2021). These synthetic outputs are designed to mimic real content convincingly, making them harmful threats to society (Katarya and Lal, 2020; Sareen, 2022). With the advancement of GAI, the concept of deepfakes has expanded beyond visual and auditory media to include textual content (Chong et al., 2023; Uchendu et al., 2023, 2024). Furthermore, AI-generated deepfake news, a typical form of "fake" text, naturally raises significant concerns (Lee and Shin, 2022; Guo, 2024).

Numerous studies have indicated that humans cannot effectively discern deepfake news from human-written news. For example, Kreps et al. (2022) conducted comparative experiments to assess the credibility of GPT-2 generated news compared to human-written news and found that individuals are largely incapable of distinguishing between AI- and human-generated text. Similarly, Fagni et al. (2021) discovered that GPT-2 can produce high-quality tweet texts that even expert human annotators can-

not unmask. This issue persists with the development of more advanced AI models; for instance, Bashardoust et al. (2024) found that GPT-4-generated fake news is perceived as less accurate than human-generated fake news, yet both types tend to be shared equally. Additionally, Chen and Shu (2023) tested GPT-3.5 and other AI models and found that, compared to human-written misinformation like fake news, AI-generated misinformation is harder for humans to detect, thus potentially causing more harm.

Compared to human detection, machine-based deepfake detection is the primary focus of current research. As Zellers et al. (2019) suggest, "the best way to detect neural fake news is to use a model that is also a generator". In the era of GAI, ChatGPT's ability to detect deepfakes has also enhanced the interest of many researchers. Alexander et al. (2024) tested ChatGPT's capability to detect visual deepfakes and demonstrated that GPT-4 can identify misleading visualizations with moderate accuracy without prior training. Koka et al. (2024) compared the detection efficiency of different AI models and indicated that GPT-4 has an accuracy of 96.7% in detecting deepfake news. Additionally, Sallami et al. (2024) found that GPT-4 exhibits significantly different levels of detection ability, achieving a 42% performance level in detecting human-created fake news but a much higher 97% in detecting deepfake news. As one of the generators of deepfakes, ChatGPT, especially GPT-4 model, seems to demonstrates strong performance in detecting the fake content it produces.

In the field of AI development, evaluating the outputs of AI models is also essential to ensure that these models are reliable, accurate, and meet business preferences. To this end, both human evaluators (Schuff et al., 2023; Tam et al., 2024) and machine-based evaluators (Chiang and Lee, 2023; Desmond et al., 2024) are tasked with assessing AI-generated outputs to test and measure how well AI models perform in real-world situations. This evaluation process can be considered deepfake detection, as it often involves distinguishing between AI-generated and human-generated content (Zhu et al., 2024). In such cases, a lower accuracy in distinguishing indicates a higher performance

of the AI model in mimicking human.

In our setting, we use human detection as the main task and also incorporate the results of AI-based detection as advice to assist human subjects. This chain-style human-AI collaboration has been developed and proven to be efficient in the process of AI model development (Wiegreffe et al., 2021; Li et al., 2023; Chu et al., 2024). On the other hand, the news texts used in our deepfake detection task are compositions of human-written and AI-generated news. This differs from most of the previously mentioned studies, where the texts are totally generated by AI or totally written by humans, without any composition. We believe that, instead of the binary classification (totally AI-generated or totally human-written), determining the specific proportion (e.g., the proportion of human-written content in combined news) may offer a more precise evaluation of human detection capabilities. Therefore, based on the above two points, we believe that deepfake detection is meaningful not only for our primary focus on AI reliance but also provides valuable insights for AI model development.

2.2 Previous Work on AI Reliance

Most studies on reliance on AI utilize systems embedded with AI algorithms. A common phenomenon observed across various fields is the relatively lower reliance on algorithms, a concept termed "algorithm aversion" by Dietvorst et al. (2015). In these studies, humans are often introduced as comparative elements. For example, research has shown that experts tend to rely less on AI algorithms compared to non-expert humans (Reverberi et al., 2022; Agarwal et al., 2023). Similar phenomena are also observed when comparing reliance on algorithms with that on human peers (Gaube et al., 2021; Mesbah et al., 2021). In contrast, Logg et al. (2019) found that people rely more on algorithms and exhibit "algorithm appreciation." Additionally, Dietvorst et al. (2018) discovered that participants tend to rely more on algorithms when they are allowed to adjust the algorithm's advice slightly.

The mixed outcomes (e.g., "algorithm aversion" vs. "algorithm appreciation") in studies on reliance on AI algorithms have prompted further investigation into the potential reasons behind. Castelo et al. (2019) indicated that "algorithm aversion" is influenced by the task's perceived objectivity and the algorithm's human likeness. Specifically, reliance on algorithm is lower for more subjective tasks compared to objective tasks, and increasing the algorithm's effective human likeness can eliminate participants' algorithm aversion when facing subjective tasks. Tse et al. (2024) further found that the performance of algorithms did not significantly affect over-reliance on algorithms when participants were given the freedom to submit any number after receiving the algorithm's advice. Maggioni and Rossignoli (2023) also suggested that "algorithm aversion" for robots could be diminished by implementing a verbal interaction between subjects and robots.

However, with the development of Large Language Models (LLMs) and the increasing popularity of GAI, there is a growing trend of "GAI appreciation" or "AI reliance" (Samarasinghe and Prasangani, 2023; Kaur, 2023; Klingbeil et al., 2024; Kumar et al., 2024), while observations of "GAI aversion" have become rare. Additionally, the effects of advice from humans versus GAI have been investigated. For instance, Leib et al. (2024) found that AI-generated advice can successfully corrupt people, but the effects of AI-generated and human-written advice are similar. Shekar et al. (2024) asked participants to evaluate medical responses written by medical doctors or generated by AI, and they found that participants could not effectively distinguish between AI-generated and doctors' responses and demonstrated a preference for AI-generated responses. Similarly, Kuosmanen (2024) examined life-related advice and found that participants could not identify AI-generated advice, and they preferred AI advice over human advice.

There are four fundamental innovations in this study. Firstly, Previous studies on GAI reliance using the JAS framework are limited and are mostly observed in research

on "algorithm aversion". In our setting, participants are asked to assess the proportion of human-written content in synthetic news, which allows us to collect numerical responses from participants and advice from AI. This enables the application of the classical JAS framework to quantify the degree of reliance using the WOA method. Secondly, we also developed a new method of DOR. Such a decomposition of reliance has never been seen in previous works. Thirdly, the deepfake detection task in our experiment is intrinsically linked to the AI advisor. Specifically, the synthetic news is generated by OpenAI's GPT-2 model, and the AI advice is provided by OpenAI's GPT-4 model. As mentioned in the previous subsection, this type of collaboration in evaluating LLMs' outputs is common in AI development, yet few studies examine reliance on AI within this context. Lastly, considering the reliance on GAI and the potential risks associated with the spread of deepfakes, our study creates a "rely on fire to fight fire" scenario to observe whether participants rely on AI to solve the problem generated by AI, thereby highlighting the duality of current GAI usage in real life.

3 Experimental Design and Hypotheses

3.1 Procedure

The experiment was programmed using Otree 5 (Chen et al., 2016), and the overall procedure is shown in Figure 1.

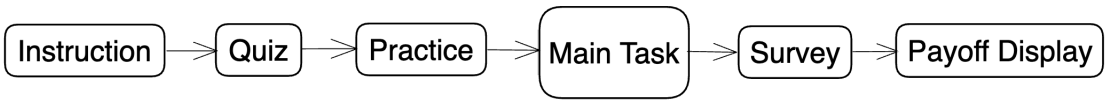

Figure 1: Overall Procedure

In the experiment, after reading through the instructions[1], each participant was asked to take a quiz (see Appendix A) to ensure they understood the rules. Then,

[1]An English translation of the instruction is provided in Online Appendix.

they practiced once and entered the main tasks. After finishing all the tasks, they were asked to complete some survey questions, and the final payoff was shown.

3.2 Main Task

The main task of the experiment consists of four stages as shown in Figure 2.

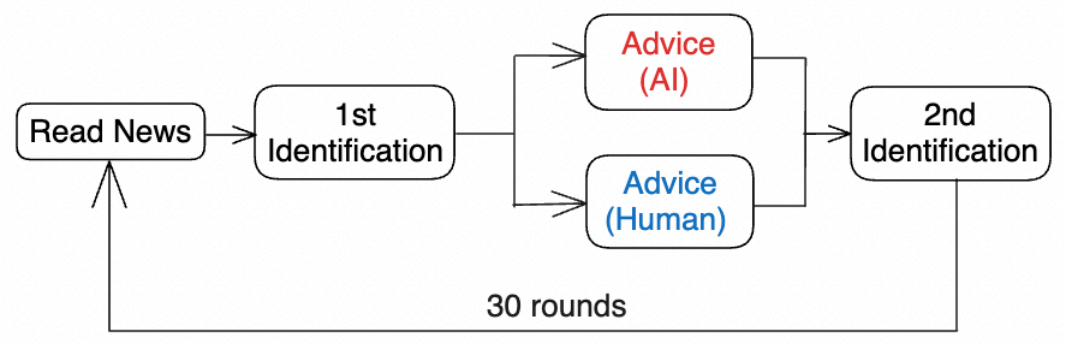

Figure 2: Main Task

There were 30 rounds of deepfake detection tasks under the framework of JAS. In each round, participants were first asked to read a piece of deepfake news and report their first identification of its "authenticity", defined as "the proportion of the human-written part in the news". Then, they were shown the advice and asked to report their second identification. We employed a between-subject design where half of the participants were displayed advice from human peers, and the advice for the remaining half were from ChatGPT's GPT-4 model.

3.2.1 Read News Stage

In the first stage of the main tasks, each participant read the news (see Figure 11 in Appendix B) without any time constraints. The news articles were Japanese deepfake news collected from an open fake news dataset[2]. We randomly selected 30 pieces of news[3] that primarily covered topics such as politics, sports, meteorology, and public

[2]https://github.com/tanreinama/japanese-fakenews-dataset?tab=readme-ov-file
[3]The original text in Japanese is available upon request.

safety. The news came in three types, as described in Table 1, where the "Length" refers to the number of the characters.

Table 1: News Materials

Type	Count	Min. Length	Max. Length	$realr$
Totally real	10	317	460	100
Totally fake	10	309	462	0
Partially fake	10	323	393	$(0, 100)$

The totally real news was written by humans and collected from Japanese wiki news[4], the totally fake news was generated by OpenAI's GPT-2 Japanese model, and the partially fake news was the composition of real and fake, where the first part of the article was human-written and the second part was AI-generated news.

The proportion of the human-written part, $realr \in [0, 100]$, considered as the degree of "authenticity", is defined as

$$realr^s = \frac{\text{the length of human-written part of the News in Round } s}{\text{the length of the News in Round } s} \times 100$$

, where $realr^s = 0$ represents totally fake news, $realr^s = 100$ represents totally real news, and $realr^s \in (0, 100)$ represents partially fake news in round s.

In the instructions, students are clearly informed about the composition style of the news materials, along with the definition of $realr$. They are also informed that $realr$ represents the degree of "authenticity" they must identity and report. The 30 pieces of news's $realr$ were assigned in a random sequence as shown in Figure 3.

[4]https://ja.wikinews.org/wiki

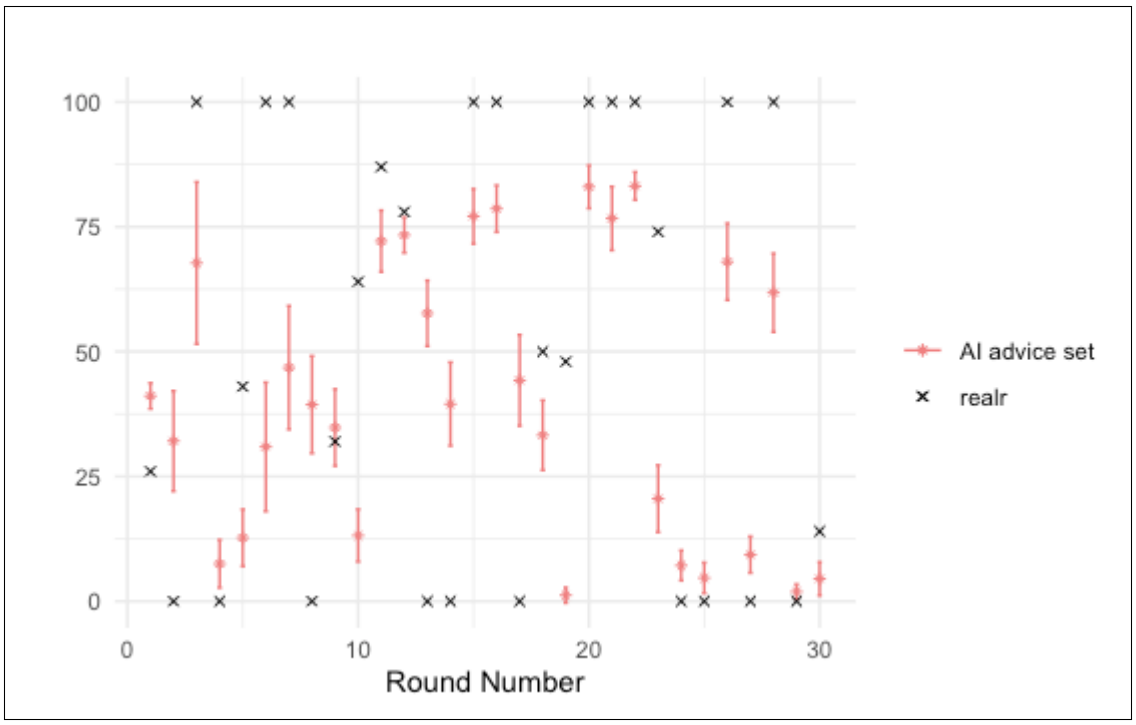

Figure 3: *realr*, the original AI advice set and Round Number

Note: The points marked with "×" represent the *realr* values in the 30-round tasks. The red points indicate the original AI advice (24 data points in each round) generated before the experiment, and the 95% CI.

3.2.2 First Identification Stage

After each participant had read the news, they were asked to report a number between 0 and 100 to represent their first identification with a slider (see Figure 12 in Appendix B). This stage was displayed on a separate screen to prevent participants from adjusting the slider while reading. This allows us to separate the time spent reading the news from their first identifications. Each participant was required to submit their first identification, *Initial Response*, in this stage; otherwise, they could not proceed to the next page.

3.2.3 Advice Display Stage

In the Advice Display Stage, we divided all participants into two distinct groups: the **AI group** and the **Human group**, and each participants was displayed a pieces of advice either from AI or Human peers. In detail, in the **AI group**, the advice was one response randomly selected from 24 responses generated by ChatGPT's GPT-4 model. Conversely, in the **Human group**, the advice was randomly selected from another participant's first identification (*Initial Response*) within the same group.

The AI's advice set was generated by GPT-4 before the experiment, using the same prompt as the task descriptions provided in the instructions to participants. The prompt was as follows:

> – *We will now send you some Japanese news. Please identify how real it is and report your belief in its authenticity as an integer from 0 to 100, with 0 representing totally fake and 100 representing totally real news.*
> – *Do not say anything else about the result of your identification.*

We added the final sentence of the prompt to limit ChatGPT's response to a number. For each piece of news, we repeated this process 24 times to generate 24 distinct responses and the original AI advice set is shown in Figure 3 with a 95% confidence interval (CI). On the experimental screen (see Figures 13 and 14 in Appendix B), the advice was displayed not merely as a number following its source name but as a screenshot of the response, which reinforce participants' belief that the advice was genuinely generated from ChatGPT. In the **AI group**, since the screenshot of the advice included the news article, we similarly represented news articles in the Advice Display Stage for the **Human group**. To prevent participants from spending excessive time rereading the news or lingering at this stage, we introduced a 10-second time constraint. Nevertheless, participants could proceed to the next page earlier by clicking the "next" button on the screen.

Note that in this stage of each round, we randomly selected a advice for each

participant. That is, the advice for the same news may differ among participants in each round. We introduced this to prevent the "spotlight effect" (Gilovich et al., 2000), which refers to the tendency for individuals to overestimate the extent to which their actions or appearance are noticed by others. By ensuring that each participant encounters a unique set of advice, bias related to this effect in the **Human group** could be eliminated.

3.2.4 Second Identification Stage

The final stage provided participants with an opportunity to adjust their initial response. At this stage, participants were required to submit their second identification (*Final Response*), which was not obligated to align with their *Initial Response*. To remind participants of their previous response and the advice provided, their *Initial Response* and the *advice* in that round were distinctly marked on the slider with different colors (see Figure 15 in Appendix B).

3.3 Survey Questions About Prior Beliefs

As well as demographic information-related questions (see Appendix C), participants' prior beliefs were obtained using the following two questions.

1. *Have you heard about ChatGPT?*

2. *In today's experiment, specifically in the task of "assessing News' authenticity," which do you think can provide more accurate responses?*

The second question offered three response options: "Generative AI," "Human," or "Not sure." We designed this to observe the consistency between their prior beliefs and their actual experimental choices.

3.4 Final Payoff

The participants' final payoff consists of a fixed participation fee and an additional performance-based payoff. Specifically, each participant received a participation fee of 500 JPY, and the additional payoff , π, was calculated based on the accuracy of one randomly selected response from all their responses throughout the experiment (a total of 30 rounds $\times$ 2 responses = 60 responses), determined using the following quadratic equation:

$$\pi = \max\{0,\ 2300 - 0.3 \times (R^{rd} - realr^{rd})^2\} \text{ JPY},$$

where R^{rd} is the randomly selected response, and $realr^{rd}$ denotes the corresponding proportion of the Human-written part of the news in the selected round, after rounding.

According to the payoff calculation formula, a participant would secure a positive additional payoff as long as the difference between the R^{rd} and the corresponding $realr^{rd}$ is less than 88. If R^{rd} exactly matches the $realr^{rd}$, the participant will obtain the maximum additional payoff of 2300 JPY.

3.5 Materials and Summary

The experiment was conducted in the laboratory at the Institute of Social and Economic Research ($ISER$) at Osaka University on November 7 and 9, 2023, and October 28 and 29, 2024. We recruited 87 participants who were students at Osaka University registered in the ORSEE (Greiner, 2015) database of $ISER$. All participants were native Japanese speakers, 42 out of whom were assigned to the **Human group** and 45 to the **AI group**[5]. In the final sample, 77% of the participants were male, and 62% were undergraduate students, predominantly from the following majors: 53% engineering, 24% medicine, 7% law, and 4% human science. Comparisons of demographic data at

[5]A power analysis (power= 0.8, Bonferroni-adjusted significant level = 0.025) based on the result of a pilot experiment (the effect size, $d = 0.175$) suggests that we at least need 21 participants answering 30 tasks in each group.

the 95% CI level, illustrated in Figure 4 and variable definitions presented in Table 2.

During the experiment, participants were prohibited from using any of their own electronic devices, including smartphones and tablets. Although they completed the tasks on the laboratory's computers, Internet connectivity within the experiment software was aslo disabled.

As a result, after completing the 90-minute experiment, participants in the **Human group** earned an average final payoff of 2373 JPY, while those in the **AI group** earned 2429 JPY. As there were 30 round tasks for each participant, the final sample size was 2610 (1350 in the **AI group** and 1260 in the **Human group**).

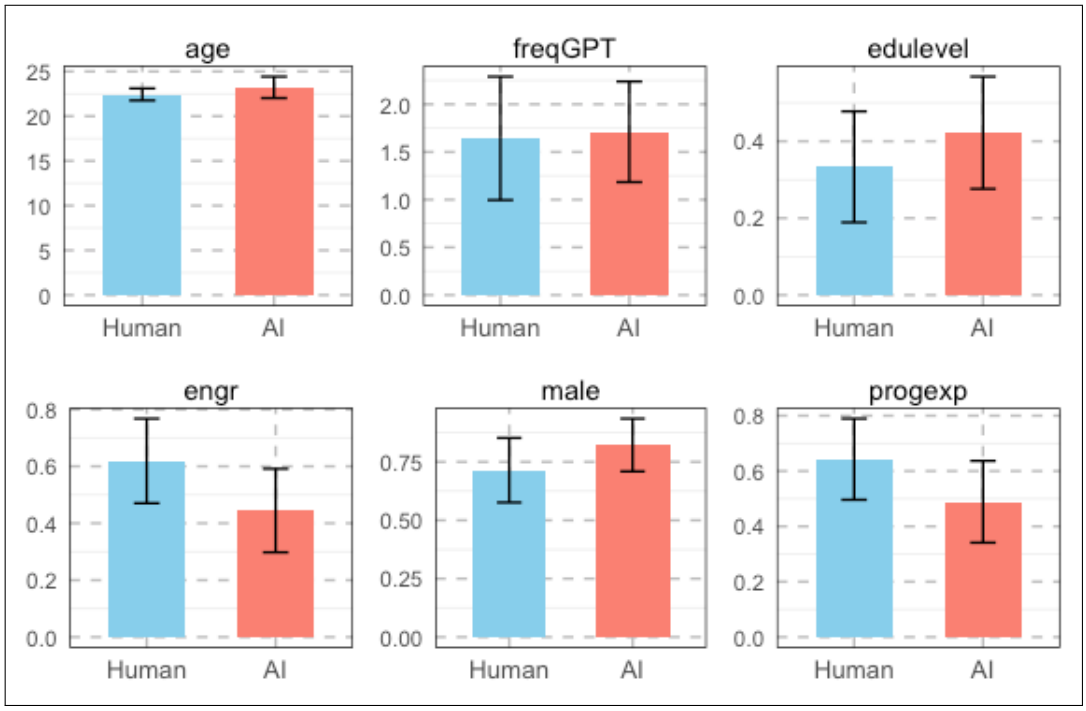

Figure 4: Demographic Comparisons

3.6 Hypotheses

As noted in the **Literature Review** section, there is a growing trend of "AI reliance". However, there is little systematic research to clarify its potential reasons. We consider that this trend may be related to some new features of GAI compared to the

Table 2: Demographics

Var.	Definition	Min.	Max.	Avg.	S.D.
age	Participants' age number.	18	44	22.8	3.34
freqGPT	Average days per week using ChatGPT.	0	7	1.68	1.96
edulevel	Participants' education level; = 1 if graduate; = 0 if undergraduate.	0	1	0.38	0.488
engr	Participants' major; = 1 if majoring in engineering.	0	1	0.53	0.502
male	Gender; = 1 if the participant is male.	0	1	0.77	0.423
progexp	Programming experience; = 1 if the participant has programming experience.	0	1	0.56	0.499

AI algorithms studied previously. As Castelo et al. (2019) suggested, the degree of reliance on AI may be affected by the task's perceived objectivity and the AI system's human likeness, i.e., higher perceived objectivity of the task and greater human likeness of the AI system lead to increased reliance on that AI system.

In this study, we employ ChatGPT as the AI advisor. As one of the most popular GAI products, ChatGPT's powerful verbal and interactive capabilities, along with its creativity, have led to its perception as possessing a high level of human likeness compared to other AI algorithms (Tseng et al., 2023; Gao et al., 2023; Azaria, 2023; Cai et al., 2024). Moreover, although "detecting fake news" is often considered a subjective task (Giglietto et al., 2016; Waisbord, 2018), we propose that "detecting deepfake news" should be a relatively objective task. This is because deepfake news is highly constructed and has an objective standard (e.g., the proportion of human-written parts), whereas real-world fake news, typically written by humans, is too complex for AI systems to handle effectively.

Therefore, based on ChatGPT's human likeness and the objectivity of deepfake detection, we propose the following hypothesis.

H1: The reliance level on the external advice source is higher in **AI group** than in

Human group.

As noted, the news materials used in the tasks consisted of three types of news: totally real, partially fake, and totally fake, with AI-generated content ranging from 0% to 100%. Since ambiguous content is more difficult for individuals to detect than extreme content (Friggeri et al., 2014; Lewandowsky et al., 2017), we believe that in deepfake news detection, determining a specific proportion of human-written part ($realr \in [0, 100]$) is much harder than simply judging whether the news is fake or real ($realr \in \{0, 100\}$). Consequently, when faced with this more difficult task, participants may be more inclined to rely on external assistance.

Therefore, we hypothesize that the degree of reliance on ChatGPT becomes higher in more challenging tasks. This leads us to the second hypothesis:

H2: The reliance level increases when the news is partially fake compared to when it is totally fake or totally real.

4 Weight of Advice

4.1 Definition of WOA

We measured the degree of "reliance on external advice source" (Castelo et al., 2019; Schemmer et al., 2022) by the "weight of advice" (Önkal et al., 2009), which is defined for participant i in the task of round s, as follows.

$$WOA_i^s = \frac{Final\ Response_i^s - Initial\ Response_i^s}{Advice_i^s - Initial\ Response_i^s}$$

As our first analysis method, this quantification approach provides a continuous outcome on a scale from 0 (completely ignoring the advice) to 1 (completely relying on the advice). Additionally, a WOA greater than 0.5 indicates that the participant's final response is closer to the external advice than to their own initial response, representing

a relatively higher degree of reliance on external advice. Conversely, a WOA less than 0.5 signifies a lower degree of reliance on external advice. We then dropped 116 observations where WOA was undefined (i.e, when the initial response was equal to the advice given), resulting in an adjusted sample size of 2494 (1291 in the **AI group** and 1203 in the **Human group**).

Our following process of WOA analysis is as follows. We firstly compared the average WOA between the **AI group** and **Human group**. Then, we employed regression tools to investigate the effects of advice quality, news type, and prior beliefs on WOA. In all of the analyses, standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant.

4.2 WOA Comparisons

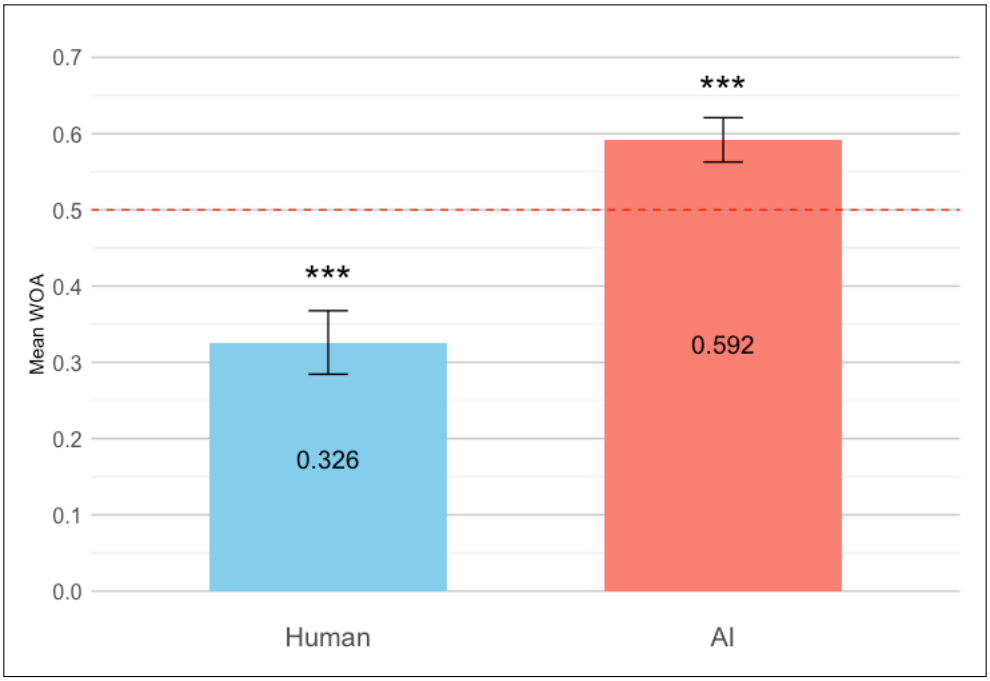

Figure 5: The Average WOA.

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The p values were calculated based on a single-sample t-test. The average WOA were compared against the 0.5 level, which is halfway between the advice and the initial response.

Figure 5 shows the average WOA with a 95% CI for the two groups. The average WOA was 0.326 in the **Human group** and 0.592 in the **AI group**, with the WOA in the **AI group** being significantly higher than that in the **Human group**[6] (Mann-Whitney U test, $p < 2.2 \times 10^{-16}$). Additionally, both groups' WOA are significantly different from 0.5.

These results show that in the **Human group**, participants exhibited a low level of reliance on external human peers' advice and tended to stick with their initial identification. In contrast, participants in the **AI group** demonstrated a high level of reliance on external ChatGPT advice and were more inclined to follow AI-generated suggestions. Overall, the degree of reliance on ChatGPT was significantly higher than that on human peers. Therefore, our first hypothesis, **H1**, is supported.

4.3 Advice Quality

We defined the "advice quality" ($advq$) for participant i in round s as "the accuracy of the advice", as follows:

$$advq_i^s = 1 - \frac{|Advice_i^s - realr^s|}{100}$$

The $advq$ is calculated as "1-the normalized absolute error". Thus, an $advq$ of 1 indicates that the advice perfectly matches the correct answer, $realr$. The average advice quality are shown in Figure 6 with a 95% CI.

Overall, AI advice exhibited an average quality score of 0.718, while human peers' advice had an average quality score of 0.716. The difference between the two groups was not statistically significant (two-sample t-test, $p = 0.9094$), indicating that AI advice can not provide greater assistance to participants than advice from human peers.

[6]The minimum detectable effect size is 0.126 based on a power calculation (power= 0.8; Bonferroni-adjusted significant level=0.025).

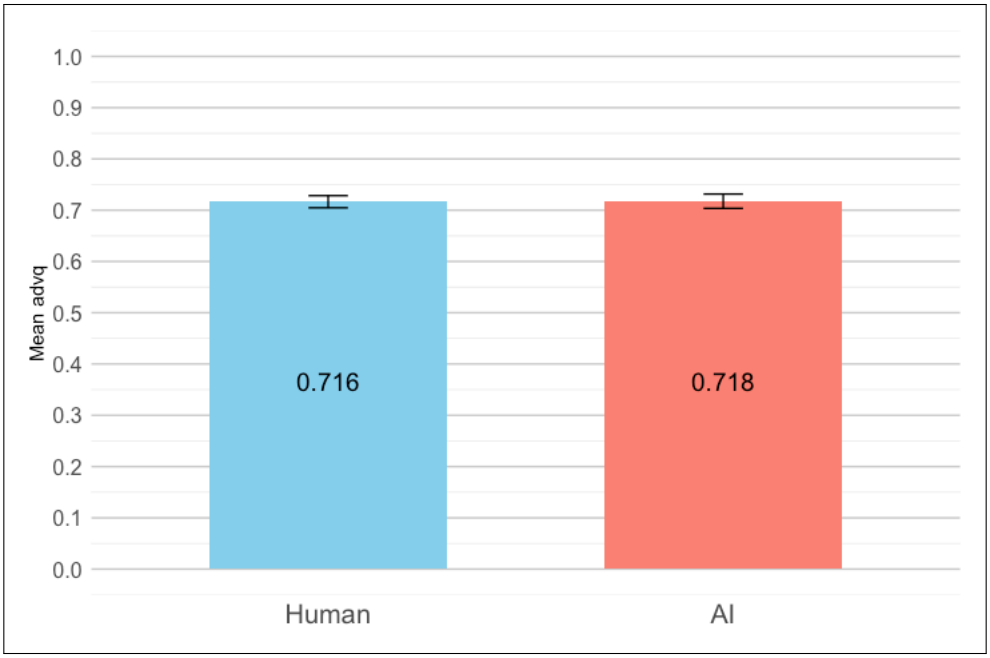

Figure 6: The Average Advice Quality

We then regressed WOA on a treatment dummy ($inAI = 0$ if **Human group** and $inAI = 1$ if **AI group**) and $advq$, along with some control variables which are related to the time spent on the task, by using OLS regression model with robust standard errors clustered by individual participants. The results are shown in Table 3.

Table 3: Source, Time, and Advice Quality

Var.	Estimate	S.E.	t value	Pr(>|t|)	Sig.level
inAI	**0.254**	**0.043**	**3.881**	**0.000**	***
advq	**0.076**	**0.043**	**1.750**	**0.080**	+
aveRead	−0.175	0.312	−0.561	0.575	
*timeidt*1	−0.004	0.005	−0.832	0.406	
*timeidt*2	−0.001	0.001	−0.232	0.218	
roundnum	−0.0001	0.002	−0.080	0.936	
Constant	0.323	0.083	3.881	0.000	***

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *aveRead* denotes the time a participant spends reading every character of the news article, *timeidt*1 and *timeidt*2 are the time participants spend on the first and second identifications in each round, respectively, *roundnum* is the number of rounds.

There is a marginally significant effect of $advq$ on WOA, suggesting that during

the advice utilization process, besides the advice source, participants did not ignore the specific quality of the advice. In other words, at least at a 90% significance level, participants to some extent carefully judged and considered whether the numerical values in the advice were superior to their initial identifications when assessing *realr*. They simultaneously considered both the advice source and the quality of the advice itself. Such a conclusion can also be confirmed in other regressions presented in Table 4 and Table 5.

4.4 News Type

Along with the news types described in Table 1, we defined the news types in our study based on *realr*, a continuous variable where higher *realr* indicate more "real" news. This allows us to conduct OLS regression analysis in Table 4.

The results suggest that the news type does not significantly impact participants' reliance on external advice. During the advice utilization process, participants did not distinguish between totally fake, totally real, and partially fake news. Whether the deepfake news was extreme (totally real & totally fake), which is easy to detect, or ambiguous (partially fake), which is difficult to detect, was completely disregarded when they received advice. Therefore, our second hypothesis, **H2**, is rejected.

4.5 Prior Belief

Since all participants answered that they have heard about ChatGPT, we investigated the effect of their prior beliefs by analyzing the responses to the second question in Section 3.3. In this question, students were asked to choose between GAI and humans regarding which could perform better in the deepfake detection task before see their final payoff. The results are shown in Figure 7.

Table 4: *WOA* and News Type

	WOA						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
inAI	**0.266*****	**0.266*****	**0.263*****	**0.291*****	**0.266*****	**0.287*****	**0.255*****
	(0.042)	(0.042)	(0.042)	(0.043)	(0.042)	(0.039)	(0.047)
realr	0.0002	0.0002	0.001*	0.0005			
	(0.0003)	(0.0003)	(0.001)	(0.001)			
advq		**0.083***	**0.169***	**0.080**$^{+}$	**0.084***	**0.080**$^{+}$	**0.082***
		(0.042)	(0.066)	(0.041)	(0.042)	(0.041)	(0.041)
$realr \times advq$			-0.002^{+}				
			(0.001)				
$inAI \times realr$				-0.0005			
				(0.001)			
isreal					0.028	0.061	0.028
					(0.034)	(0.057)	(0.034)
$isfake$					-0.012	-0.012	-0.029
					(0.024)	(0.024)	(0.036)
$inAI \times isreal$						-0.064	
						(0.059)	
$inAI \times isfake$							-0.034
							(0.042)
Constant	0.317***	0.255***	0.193***	0.245***	0.260***	0.252***	0.267***
	(0.027)	(0.038)	(0.048)	(0.042)	(0.036)	(0.037)	(0.037)
Adjusted R^2	0.041	0.041	0.041	0.041	0.041	0.041	0.041
Number of cluster	87	87	87	87	87	87	87
Number of Obs.	2494	2494	2494	2494	2494	2494	2494

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. $isfake$ and *isreal* are two dummies that take the value 1 if the news is totally fake or totally real, respectively, and 0 otherwise.

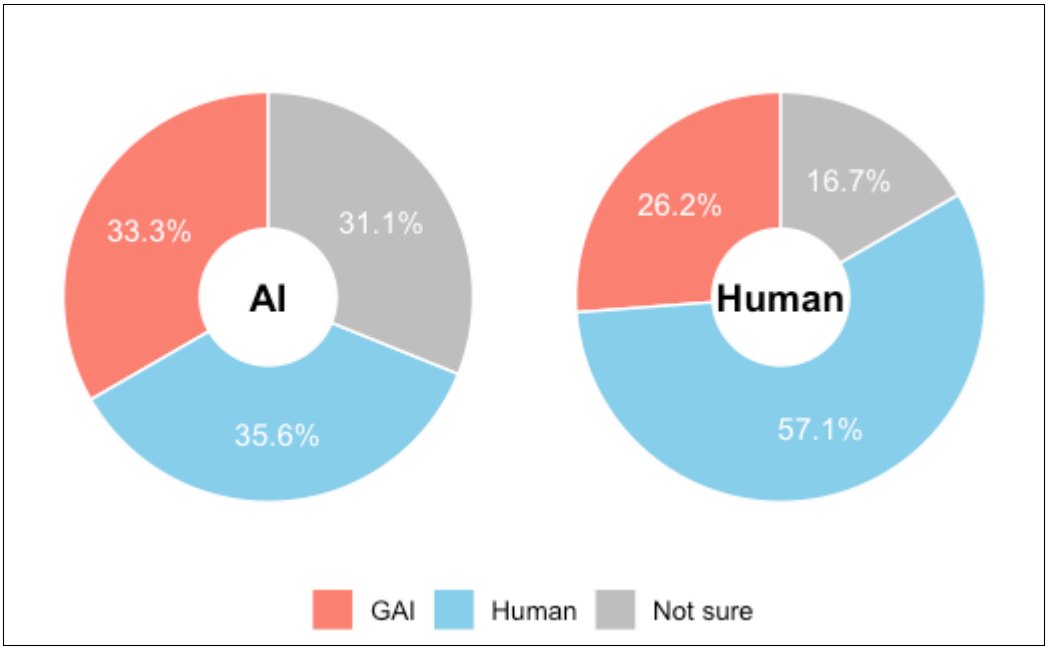

Figure 7: AI or Human Preference

Note: The left plot shows the distribution for the **AI group**, and the right plot shows the distribution for the **Human group**. The numbers on the pie charts represent the proportion of participants who chose each corresponding option. There is no significant difference between the two groups (Fisher's Exact test, $p = 0.1176$).

The question investigated participants' prior preference for AI or Humans, specifically in the deepfake news detection task. Overall, in the entire sample, about 46% of the participants believe humans will perform better, 29.9% believe AI perform better, and the remaining 24.1% indicated that they are not sure about their answer. As the experiment uses a between-subjects design, to uniformly analyze this prior belief, we used a dummy variable, $CPBS$, to represent "the consistency between prior beliefs and the actual advice source" for participant i. The definition is as follows:

$$CPBS_i = \begin{cases} 1 & \text{if } i \text{ in AI (Human) group trusted AI (Human) perform better} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, a $CPBS_i = 1$ indicates that participant i received advice from the source they prefer, and $CPBS_i = 0$ indicates that they believe the advice source may not offer

them a good advice. Then, by using $CPBS$ along with other demographic information, we investigated the relationship between reliance and prior beliefs by running OLS regressions on WOA, as shown in Table 5.

The sign of $CPBS$ shows that participants are more likely to rely on the external advice when the source is consistent with their prior belief. Receiving advice from a preferred source increases their reliance on that advice. Specifically, participants who received ChatGPT's advice and believed that GAI outperforms humans were more likely to rely on AI advice.

The influence of demographics is also worth noting. Although the "frequency of using ChatGPT" ($freqGPT$), "gender" ($male$), "age" (age), and "education level" ($edulevel$) have no significant effect on WOA, "having programming experience" ($progexp$) has a negative effect, while "majoring in engineering" ($engr$) has a positive effect on WOA. However, these two significant effects do not differ significantly between the **AI group** and **Human group**.

Notably, individuals who have programming experience may place greater trust in their own initial identifications and rely less on external advice, regardless of whether the source is from ChatGPT or human peers. In contrast, those majoring in engineering exhibit the opposite behavior, still regardless of the advice source. Since individuals with programming experience and those majoring in engineering are more likely to encounter the field of deepfake detection (i.e., machine learning), we considered that this result may stem from participants' perception of the deepfake detection task itself, rather than the advice source, specifically ChatGPT.

Table 5: WOA and Prior Belief

	WOA					
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	**0.301*****	**0.235*****	**0.300*****	**0.303*****	**0.304*****	**0.277*****
	(0.039)	(0.056)	(0.039)	(0.053)	(0.048)	(0.053)
advq	**0.073**$^{+}$	**0.077**$^{+}$	**0.130****	**0.073**$^{+}$	**0.073**$^{+}$	**0.072**$^{+}$
	(0.039)	(0.039)	(0.049)	(0.039)	(0.039)	(0.039)
CPBS	**0.082***	0.013	**0.182****	**0.082***	**0.082***	**0.079***
	(0.038)	(0.047)	(0.064)	(0.038)	(0.040)	(0.038)
freqGPT	−0.008	−0.010	−0.008	−0.008	−0.008	−0.009
	(0.011)	(0.011)	(0.011)	(0.014)	(0.011)	(0.011)
male	−0.004	0.012	−0.005	−0.005	−0.004	−0.004
	(0.050)	(0.048)	(0.050)	(0.051)	(0.050)	(0.050)
progexp	**−0.115****	**−0.116****	**−0.115****	**−0.115****	**−0.115****	**−0.138***
	(0.043)	(0.043)	(0.043)	(0.044)	(0.044)	(0.056)
age	0.006	0.004	0.006	0.006	0.006	0.006
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
edulevel	−0.040	−0.029	−0.040	−0.041	−0.040	−0.041
	(0.044)	(0.044)	(0.044)	(0.044)	(0.045)	(0.044)
engr	**0.193*****	**0.181*****	**0.193*****	**0.193*****	**0.196*****	**0.195*****
	(0.042)	(0.044)	(0.042)	(0.041)	(0.049)	(0.042)
$inAI \times CPBS$		**0.139**$^{+}$				
		(0.075)				
$advq \times CPBS$			**−0.140**$^{+}$			
			(0.079)			
$inAI \times freqGPT$				−0.001		
				(0.021)		
$inAI \times engr$					−0.005	
					(0.076)	
$inAI \times progexp$						0.042
						(0.075)
Constant	0.078	0.148	0.038	0.075	0.077	0.085
	(0.103)	(0.107)	(0.108)	(0.102)	(0.104)	(0.103)
Adjusted R^2	0.060	0.062	0.061	0.060	0.060	0.060
Number of cluster	87	87	87	87	87	87
Number of Obs.	2494	2494	2494	2494	2494	2494

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Definitions of demographic variables are shown in Table 1.

5 Decomposition of Reliance

5.1 From WOA to DOR

The first method, WOA, has been broadly used in lots of advice-taking studies. However, there are still some issues. For example, when the initial response happens to be close or equal to the advice given, the outcome of WOA becomes larger than one or even infinity, which makes it hard to reflect the actual process of belief updates in advice-taking in real-world contexts. In the literature, most studies typically clip the outcome to have a maximum magnitude of one (Yaniv, 2004; Gino, 2008; Tinghu et al., 2018), resulting in a misrepresented degree of reliance.

In this study, we chose not to clip the WOA outcomes but instead excluded instances where the outcome was infinity to maintain data integrity. This resulted in approximately 4.45% of the total sample being unused. Upon reviewing the 116 excluded samples, we found that in 17 cases, the final response differed from the initial response. This indicates that, in these instances, participants adjusted their initial identification after receiving the advice, even when their initial identification fully matched the advice. The fact that participants made adjustments despite the advice aligning with their initial responses raises our curiosity. Meanwhile, the remaining 99 cases, where the *Initial Response*, *Advice*, and *Final Response* are all equal, prompt the question of whether this lack of adjustment is due to participants' willingness to fully adhere to the advice (indicating complete reliance on it) or an unwillingness to alter their initial identification (indicating strong confidence in their own judgment).

This consideration prompted us to develop a second method, DOR, to elucidate the reliance, allowing us to utilize the entire sample without making any adjustments such as clipping or dropping.

Specifically, the DOR method decomposes reliance into two stages of the advice utilization. This approach follows the "activation-integration" model proposed by Vo-

drahalli et al. (2022). In the first (activation) stage, a participant decides whether to use the advice. In the second (integration) stage, participants integrate their own experience and knowledge to determine the extent to which they will use the advice. In the analysis, they employed two mixed-effects models and concluded that the source of advice influences the activation stage but not the integration stage.

In the following parts of this section, we employed the activation-integration model to further dissect the reliance mechanism, utilizing the Heckman correction method (Heckman, 1974; Heckman and RajBhandary, 1979) for the analysis. We discussed the activation and integration stages separately. In the activation stage, we examined the factors that had a significant effect in the WOA analysis, as shown in the previous section. Then, in the integration stage, we incorporated the Heckman correction method into the two-stage model and conducted a comprehensive analysis of advice utilization in the integration stage. In all of the analyses, standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant.

5.2 Activation Stage

In the analysis of Vodrahalli et al. (2022), they defined a participant as "activated" for a given task if they change their initial judgments by at least a threshold (3.5% of the length of the slider they used) amount after receiving advice. In our analysis, we improved it by reducing the threshold to zero. Therefore, all the status of observations in the activation stage can be defined as follows.

$$Acti_i^s = \begin{cases} 1 & \text{if } Final\ Response_i^s \neq Initial\ Response_i^s \\ 0 & \text{if } Final\ Response_i^s = Initial\ Response_i^s \end{cases}$$

Specifically, a participant is considered "activated" if they adjusted their initial

response after receiving a advice and "not activated" if they maintained their initial response. In our study, 2116 out of the 2610 total samples were activated, resulting in an overall activation rate of 81.1%. The average *Acti* across different groups are depicted in Figure 8.

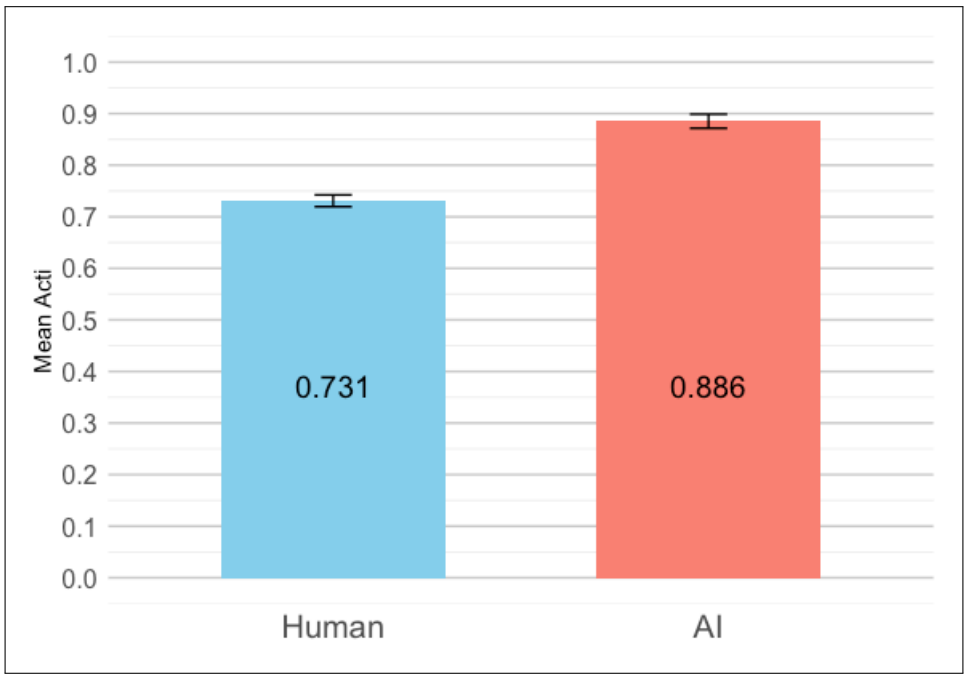

Figure 8: Activation Rate

The average *Acti* in the **Human group** is 0.731, indicating an average of activation rate of 73.1%, which is significantly lower than the 88.6% observed in the **AI group** (Chi-Squared test,$\chi^2 = 100.03, p < 2.2 \times 10^{-16}$, Cramér's V$\approx$ 0.197). Results of Probit regressions are shown in Table 6, where we added the "gap between the advice and initial response" as a new control variable, denoted as *advGap* and calculated as $|Advice - Initial\ Response|$.

As indicated by the results presented, participants who received a advice from ChatGPT rather than from their human peers tended to be activated more frequently. In addition, participants who received advice from a source they perceived as more effective for the task showed a higher likelihood of activation. Furthermore, our analysis revealed that the greater the gap between the advice and the initial identification (*advGap*), the

Table 6: Probit Regressions of *Acti*

	Acti			
	(1)	(2)	(3)	(4)
inAI	**0.675*****	**0.675*****	**0.689*****	**0.692*****
	(0.164)	(0.164)	(0.151)	(0.162)
advq	0.081	0.081		0.089
	(0.152)	(0.152)		(0.153)
aveRead	0.025	0.021		0.098
	(1.360)	(1.363)		(1.225)
realr	0.00004			0.0001
	(0.001)			(0.001)
isreal		−0.007		
		(0.066)		
isfake		−0.023		
		(0.064)		
CPBS	**0.246**$^{+}$	**0.246**$^{+}$	0.233	0.230
	(0.149)	(0.149)	(0.165)	(0.156)
advGap	**0.020*****	**0.020*****	**0.020*****	**0.020*****
	(0.003)	(0.003)	(0.003)	(0.003)
freqGPT			0.009	0.009
			(0.060)	(0.059)
progexp			−0.154	−0.152
			(0.165)	(0.165)
edulevel			−0.032	−0.030
			(0.167)	(0.162)
engr			0.206	0.208
			(0.154)	(0.155)
Constant	0.003	0.016	0.047	−0.042
	(0.293)	(0.296)	(0.140)	(0.278)
AIC	2288	2290	2282	2287
Number of cluster	87	87	87	87
Number of Obs.	2610	2610	2610	2610

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant.

higher the probability of participant activation. Consequently, we identified the three core factors significantly influencing whether an individual is activated: "the source of the advice" ($inAI$), "the consistency between prior beliefs and prior beliefs" ($CPBS$), and "the gap between the advice and their initial response" ($advGap$).

An interesting finding is that the quality of advice ($advq$) had no significant effect on activation, whereas it was significantly positive when tested with the WOA method. This suggests that in the first stage of advice utilization (activation), the decision to utilize the advice is primarily influenced by the source of the advice rather than its quality. This implies that participants did not place significant emphasis on assessing the quality of the advice when deciding whether to use it initially.

5.3 Integration Stage

In the second stage of advice processing (integration), our focus shifts to examining the extent to which participants utilize the advice once they decide to use it. In other words, we are interested in examining the extent of advice utilization among participants who are activated. To quantify this extent of utilization, we constructed a continuous variable as follows:

$$Itgr_i^s = \begin{cases} Final\ Response_i^s - Advice_i^s & \text{if } Advice_i^s > Initial\ Response_i^s \\ |Final\ Response_i^s - Advice_i^s| & \text{if } Advice_i^s = Initial\ Response_i^s \\ Advice_i^s - Final\ Response_i^s & \text{if } Advice_i^s < Initial\ Response_i^s \end{cases},$$

which describes "the consistency with the advice" for participant i in round s. Therefore, $Itgr > 0$ indicates that a participant moved their second identification's point ($Final\ Response$) on the slider beyond the advice point ($Advice$), thereby "over-utilizing" the advice. Conversely, $Itgr = 0$ signifies that the participant adjusted the "$Final\ Response$" point on the slider to exactly match the advice point; thus,

they "totally utilized" the advice. Finally, $Itgr < 0$ suggests that the participant did not move "*Final Response*" sufficiently on the slider to reach or exceed the advice point, indicating the "under-utilization" of the advice. Examples of these scenarios are provided in Appendix D. This classification allows us to delineate the status of advice utilization in the integration stage. Table 7 shows the proportion of these three statuses among the activated sample across different groups.

Table 7: Proportion of Utilization Statuses (%)

Group	Under-utilize	Totally utilize	Over-utilize
Human	91.0	4.34	4.66
AI	76.9	**18.2**	4.90
Total	83.1	12.1	4.80

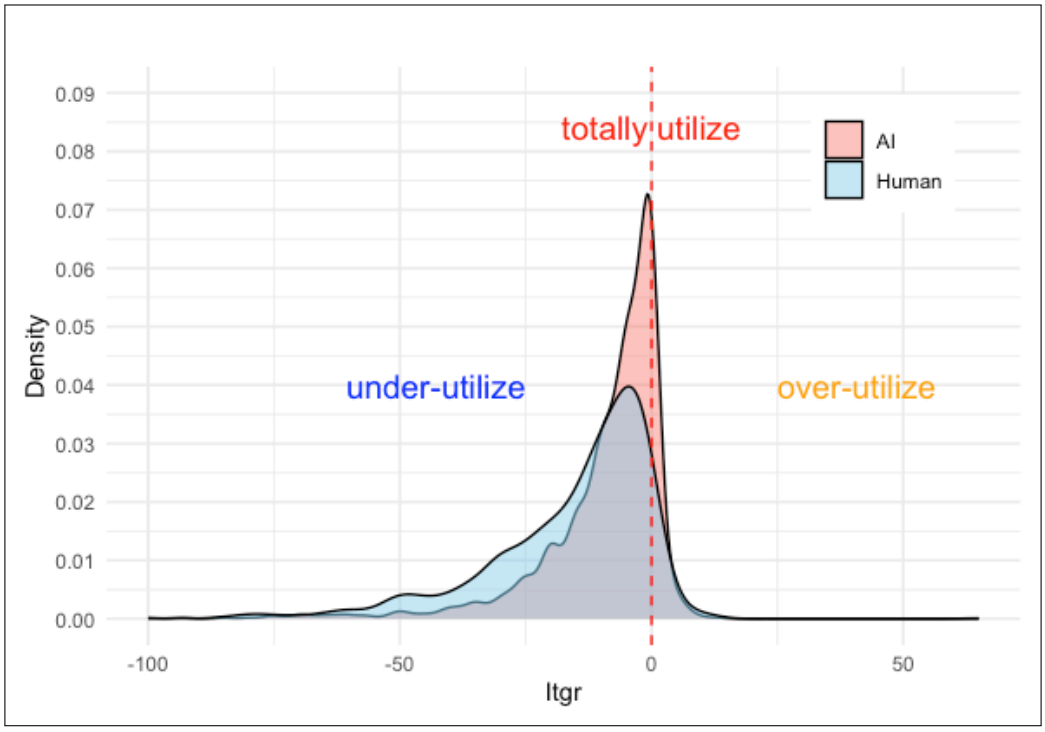

Figure 9: Distribution of Itgr

Note: The red dashed line represents the 0 value of $Itgr$, that participants "totally utilize" the advice. The distribution to the left of the red line corresponds to "under-utilize", while the distribution to the right corresponds to "over-utilize". The Mann-Whitney U test results ($p < 2.2 \times 10^{-16}$) indicate a significant difference in the distributions of $Itgr$ between the two groups.

The proportion of "totally utilize" status of ChatGPT's advice (18.2%) is about four times that of human peers' advice (4.34%)[7], but this proportion in "over-utilization" is close (4.66% vs. 4.90%), indicating that participants are more willing to fully follow ChatGPT's advice. This is also shown in the density distribution in Figure 9, where the **AI group**'s $Itgr$ distribution is more concentrated, especially at $Itgr = 0$. This finding, from another perspective, suggests the relatively excessive reliance on AI.

Here, we further applied the following Heckman correction model to estimate the effect of this two-stage model.

First Stage: *Activation* (Selection):

$$Acti = \alpha_0 + \alpha_1 \cdot inAI + \alpha_2 \cdot advGap + \alpha_3 \cdot CPBS + \alpha_4 \cdot X_1 + \varepsilon$$

Second Stage: *Integration* (Outcome):

$$Itgr = \beta_0 + \beta_1 \cdot inAI + \beta_2 \cdot CPBS + \beta_3 \cdot X_2 + \gamma \cdot imr + u$$

The equation for the **First Stage**: "*Activation*" is a probit regression model applied to the entire sample, using $inAI$, $advGap$, and $CPBS$ as independent variables. The equation for the **Second Stage**: "*Integration*" is an OLS model applied to the "activated" sample ($Acti = 1$), in which $advGap$ has been excluded as an instrumental variable. X_1 and X_2 are other control variables. ε and u are residual terms. imr denotes the "inverse mills ratio", calculated by $imr = \frac{pdf(\hat{Acti})}{cdf(\hat{Acti})}$, and it can be concluded that the selection effect exists if the coefficient of imr is significant.

We applied this method to our analysis, considering that there may be threshold levels for $inAI$, $advGap$, and $CPBS$. If these variables do not reach certain thresholds, a participant might not alter their initial identification after receiving advice (i.e., not activated). However, by employing this method in the first stage, we can include

[7]Chi-Square Test, $\chi^2 = 92.706$, $p < 2.2 \times 10^{-16}$

samples that were not activated. Subsequently, in the second stage, this model allows us to adjust the regressions conducted only on the subsample. In this way, the inability to utilize the entire sample using WOA is resolved, thereby yielding more precise analytical results.

Table 8 presents the results. First, the significance of the coefficients for *imr* suggests that selection bias does exist. Second, contrary to the predictions of Vodrahalli et al. (2022), the advice source influence participants' utilization of advice in both stages. Specifically, when the source is ChatGPT, participants tend to accept the advice more frequently, and among those who have decided to accept the advice, the level of advice utilization is also higher in the **AI group**. Third, our analysis shows that participants may not consider the specific quality of the advice (*advq*) at the first stage of utilization. Only a few decided to accept the advice and then adjust their initial identification, at which point they began to perceive the quality of the advice and consider the extent to which to adjust.

6 Discussions

6.1 Comparisons of WOA and DOR

In this study, we investigated people's relative reliance on AI in the task of deepfake detection. We employed two quantification methods: WOA and DOR, using WOA in the WOA method and *Acti* and *Itgr* in the DOR method to represent reliance. Our analysis explored how reliance levels might be influenced differently depending on the quantification method used. Table 9 provides a comparative summary across the two methods.

Overall, in the comparison between WOA and DOR, for the factors that significantly affect WOA's analysis, when we decompose the process of advice utilization, both the advice source and participants' prior beliefs still play a significant role at each stage.

Table 8: Uncorrectted OLS & Heckman Correction

	Uncorrectted OLS	Heckman Correction	
	Integration	Activation	Integration
inAI	**6.369*****	**0.675*****	**25.763*****
	(0.98)	(0.16)	(1.27)
advq	**10.890*****	0.081	**4.657****
	(1.87)	(0.15)	(1.38)
aveRead	**−12.941**$^{+}$	0.025	**−16.119***
	(7.27)	(1.31)	(6.54)
realr	−0.010	0.000	−0.002
	(0.01)	(0.00)	(0.00)
CPBS	1.247	0.246	**8.880*****
	(1.06)	(0.15)	(1.10)
freqGPT	−0.019		−0.170
	(0.30)		(0.28)
progexp	**−3.080****		**−2.891****
	(1.15)		(0.98)
engr	**4.092*****		**3.406*****
	(1.08)		(0.96)
edulevel	0.080		0.064
	(1.02)		(0.99)
advGap		**0.020*****	
		(0.00)	
imr			**75.808*****
			(4.29)
Constant	-22.114***	0.003	-54.318***
	(2.34)	(0.28)	(3.31)
Number of cluster	87	87	87
Number of Obs.	2116	2610	2116

Note: $^{+}$ $p < 0.1$, * p<0.05, ** p<0.01, *** p<0.001; Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant. *imr* is the "inverse mills ratio".

Table 9: WOA v.s. DOR

	WOA	**DOR**	
	WOA	$Acti$	$Itgr$
$inAI$	$\uparrow\uparrow\uparrow$	$\uparrow\uparrow\uparrow$	$\uparrow\uparrow\uparrow$
$advq$	$\overset{\uparrow}{\cdot}$	—	$\uparrow\uparrow$
$realr$	—	—	—
$CPBS$	$\uparrow$	$\overset{\uparrow}{\cdot}$	$\uparrow\uparrow\uparrow$
$aveRead$	—	—	$\downarrow$
$progexp$	$\downarrow\downarrow$	—	$\downarrow\downarrow$
$engr$	$\uparrow\uparrow\uparrow$	—	$\uparrow\uparrow\uparrow$

Note: "$-$" indicates no significant effect at the 0.1 level. Arrows denote significant effects, where an upward arrow ($\uparrow$) represents a positive effect and a downward arrow ($\downarrow$) represents a negative effect. The number of arrows indicates the level of significance: one arrow for $p < 0.05$, two arrows for $p < 0.01$, and three arrows for $p < 0.001$. Specifically, the dashed upward arrow ($\overset{\uparrow}{\cdot}$) signifies a positive significant effect at the $p < 0.1$ level. For each variable, the most frequent significance level across all regression models in this study is selected as its representative level.

However, advice quality ($advq$), programming experience ($progexp$), and majoring in engineering ($engr$) only influence advice utilization in the second stage.

This discontinuity in participants' different stages of advice utilization, partially satisfies the definition of the "activation-integration model", which states that "In the second (integration) stage, participants integrate their own experience and knowledge to determine the extent to which they will use the advice". Our results show that there may exist a sequential mechanism, where in the first stage, participants are activated to accept the advice only by the advice source (whether the advice source is ChatGPT or a source they already trust). Once activated, they begin to integrate their experiences, such as programming or engineering, along with assessing the quality of the advice, and then make decisions regarding the extent to which they will utilize, or in other words, rely on that advice.

6.2 Performance

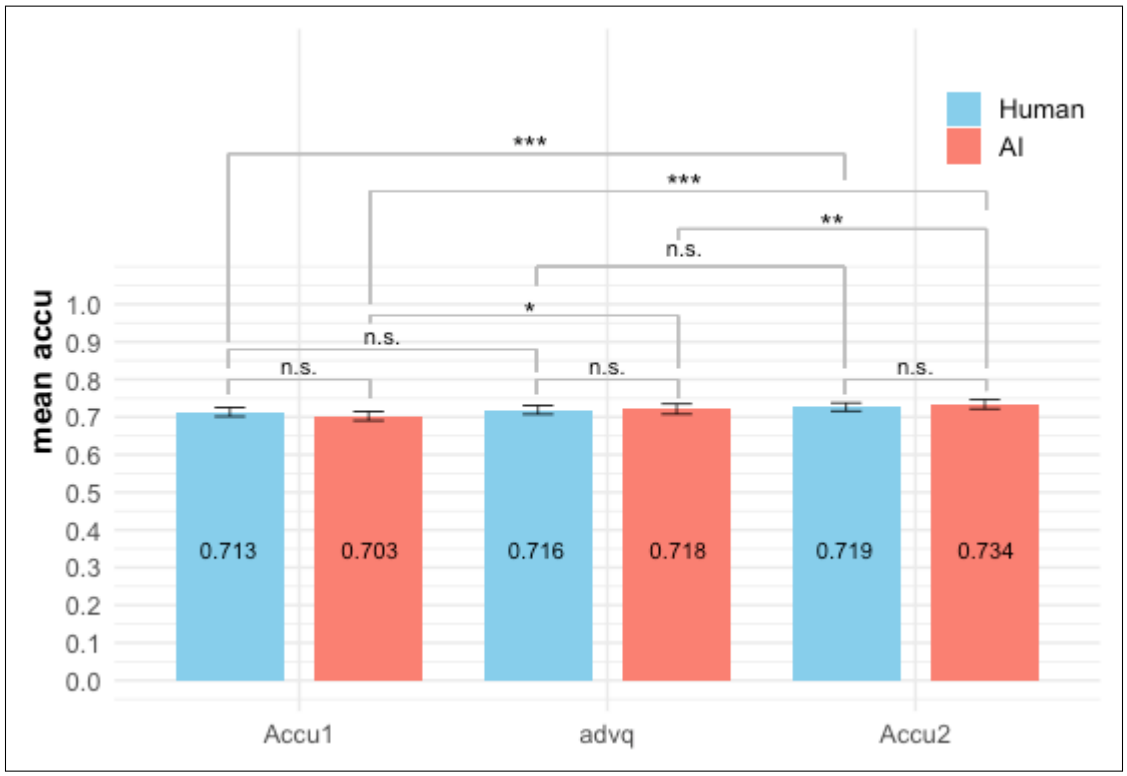

Figure 10: Accuracy

Note: We compared $Accu_1$, $advq$, $Accu_2$ within group by paired t-test and compared those between groups by two-sample t-test. $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level.

Based on the analysis of the above sections, it can be concluded that participants indeed tend to rely more on ChatGPT than on their human peers. This raises the question of whether participants can benefit from this reliance on AI. We, therefore, investigated participants' performance by calculating their accuracy as

$$[Accu_i^s]_k = 1 - \frac{|[Response_i^s]_k - realr_i^s|}{100},$$

where $k = 1, 2$ and $[Accu_i^s]_1$ is participant i's accuracy for her "*Initial Response*" ($[Response_i^s]_1$), and similar to $k = 2$.

Figure 10 presents the mean accuracy ($Accu$) and mean advice quality ($advq$) in all observations. Overall, when there is no external advice, participants' average level of

performance in the deepfake detection task is about 70.8% (71.3% in **Human group** and 70.3% in **AI group**). Although there is no significant difference between groups, all the participants behaved better after receiving external advice and adjusting their initial identification. We then test the perfomance imporvement by running OLS regressions with clustered robust standard errors, as shown in Table 10.

The sign of $inAI$ in regressions (7) $\sim$ (9) indicates that participants in the **AI group** improved their performance more than those in the **Human group**. This suggests that participants' reliance on AI is justified, as ChatGPT effectively enhanced their performance in the deepfake detection task.

The news type also influences participants' performances. The sign of $realr$ indicates that for each 1% increase in human-written characters (and thus a 1% decrease in AI-generated characters) in deepfake news, participants' accuracy decreases by approximately 0.03%. The extreme type of "totally real" ($isreal$) and "totally fake" ($isfake$) seems do not have a significant effect, or may even have a negative effect, on the accuracy of participants' initial identifications. However, they have a strong positive significant effect on participants' performance improvement. We interpret this as indicating that, for participants, extreme types of news are somewhat easier to detect. Nonetheless, participants were not confident enough to submit extreme values of 100 or 0 initially, so they chose to submit less extreme values at first and wait for advice. Subsequently, when the advice also provided a relatively extreme value, their initial suspect were to some extent confirmed, leading them to submit previously somewhat risky extreme numbers during the second identification process. This may reveal the psychological mechanism by which participants balance their own judgments with external advice.

Table 10: OLS Results of Accuracy

	$Accu_1$			$Accu_2$			$Accu_2 - Accu_1$		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
inAI	-0.014	-0.015	-0.021^{+}	0.004	0.004	-0.001	**0.018***	**0.018***	**0.020****
	(0.015)	(0.015)	(0.012)	(0.010)	(0.010)	(0.008)	(0.008)	(0.008)	(0.007)
advq				0.617^{***}	0.620^{***}	0.617^{***}	0.421^{***}	0.423^{***}	0.422^{***}
				(0.028)	(0.028)	(0.028)	(0.031)	(0.032)	(0.031)
realr	-0.0004^{+}		-0.0003^{+}	-0.0003^{*}		-0.0003^{*}	-0.00003		-0.00003
	(0.0002)		(0.0002)	(0.0001)		(0.0001)	(0.0001)		(0.0001)
isreal		-0.038^{**}			-0.005			**0.024****	
		(0.014)			(0.010)			(0.009)	
isfake		-0.014			0.013			**0.028*****	
		(0.015)			(0.011)			(0.007)	
aveRead	−0.113	−0.118		−0.077	−0.071		0.002	0.013	
	(0.116)	(0.115)		(0.078)	(0.077)		(0.064)	(0.063)	
*timeidt*1	0.00005	−0.0001		0.001	0.001		0.001	0.001	
	(0.001)	(0.001)		(0.001)	(0.001)		(0.001)	(0.001)	
*timeidt*2				-0.0002	-0.0002		0.0001	0.0001	
				(0.0003)	(0.0003)		(0.001)	(0.001)	
engr			**-0.051*****			**-0.038*****			0.013
			(0.015)			(0.009)			(0.010)
progexp			0.001			0.007			0.005
			(0.015)			(0.009)			(0.010)
edulevel			0.013			0.006			-0.005
			(0.014)			(0.009)			(0.009)
Constant	0.746^{***}	0.747^{***}	0.757^{***}	0.304^{***}	0.284^{***}	-0.293^{***}	-0.316^{***}	-0.297^{***}	-0.252^{***}
	(0.004)	(0.004)	(0.015)	(0.026)	(0.025)	(0.024)	(0.024)	(0.026)	(0.023)
Adjusted R^2	0.012	0.011	0.010	0.466	0.465	0.472	0.310	0.314	0.311
Number of cluster	87	87	87	87	87	87	87	87	87
Number of Obs.	2610	2610	2610	2610	2610	2610	2610	2610	2610

Note: $^{+}$ $p < 0.1$, * p<0.05, ** p<0.01, *** p<0.001. Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant.

7 Conclusions

In this paper, we conducted a laboratory experiment to explore people's reliance on ChatGPT in the task of deepfake news detection. Compared with previous studies on advice-taking using AI, this study innovates by 1) introducing the deepfake detection task into the laboratory, 2) using two methods to verify the over-reliance on AI in the deepfake detection task, and 3) uncovering a sequential mechanism in people's advice utilization process by decomposing reliance.

Our experiment revealed that participants tend to rely more AI than on their human peers to detect (AI-generated) deepfake news. While the degree of reliance does not vary with the news types or time spent, it is significantly influenced by participants' prior beliefs, advice quality, experience, and knowledge. By decomposing reliance into two stages−"Activation" (deciding whether to accept the advice) and "Integration" (determining the extent of its utilization)−we found that the advice source and the consistency with prior beliefs influence participants' reliance in both stages. Specifically, participants are more likely to accept and utilize advice to a greater extent when the advice source is ChatGPT rather than human peers, or when they believe the advice source is better for the task than the other. However, the advice quality's effect becomes significant only in the second stage, where participants also integrate their own experience and knowledge. This suggests a potential psychological mechanism of advice utilization, in which people first consider the advice source, then integrate their own knowledge and background while specifically assessing the quality of the advice. These processes do not occur simultaneously but follow a sequential order which cannot be revealed solely by the classical method of WOA.

Numerous studies have highlighted concerns about over-reliance on AI and the detrimental effects of AI-generated misinformation. However, our study suggests that relying on AI to detect AI may be advantageous. In our deepfake detection task, although the AI advisor, ChatGPT, did not provide better advice than human advisors or peers, it

nonetheless helped participants improve their performance more significantly, potentially due to their over-reliance on AI. When interacting with an AI advisor, participants tend to adjust their initial identifications more than when interacting with a human advisor, even though the quality of advice from AI and humans may be comparable. Since such a deepfake detection task are currently widely used in the development of AI models, especially LLMs, which form the basis of most GAI products, we believe our study not only contributes to a deeper understanding of AI-human interaction and the duality of GAI reliance in the GAI era, but also provides AI developers with insights from an economic experiment.

This study has several limitations that warrant further investigation. First, the deepfake news materials used in our experiment were generated by OpenAI's GPT-2 model, an older model released in 2019. We selected GPT-2 because it has been widely used in previous studies, allowing for comparability with existing research. However, significant differences exist between GPT-2 and more advanced models, such as GPT-4. Advanced models mimic human-written content more convincingly than GPT-2, making it potentially much more challenging for people to detect deepfakes generated by these models. Therefore, employing a more advanced AI model to generate deepfake news, and subsequently comparing the detection difficulty and examining the associated variance in reliance levels, could be a valuable extension. Second, we did not explore the specific content categories of the news materials. As is widely known, deepfakes can have a particularly harmful impact in politically related news, such as during major election periods in the United States. Thus, examining the effects of different types of deepfake news—such as entertainment, politics, and sports—on task difficulty and reliance levels would be a meaningful direction for future research. Third, the Heckman correction method typically requires the "exclusion restriction" to be satisfied. Generally, an exclusion restriction is necessary to produce credible estimates: there must be at least one variable that appears with a non-zero coefficient in the selection equation

but does not appear in the outcome equation, essentially serving as an instrumental variable. Unfortunately, the instrumental variable used in our model does not satisfy this restriction, as it appears to significantly influence the outcome directly in the outcome equation, which may reduce the explanatory power of our model. Despite this limitation, our findings provide valuable insights into the mechanisms underlying sequential advice utilization. Nevertheless, future research could benefit from identifying a suitable instrumental variable that satisfies the restriction or from exploring alternative models to better capture the process of sequential advice utilization. This remains a potential challenge and a promising direction for further study.

References

Agarwal, N., Moehring, A., Rajpurkar, P. and Salz, T. (2023), Combining human expertise with artificial intelligence: Experimental evidence from radiology, Technical report, National Bureau of Economic Research.

Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. and Li, H. (2019), Protecting world leaders against deep fakes., *in* 'CVPR workshops', Vol. 1, p. 38.

Alexander, J., Nanda, P., Yang, K.-C. and Sarvghad, A. (2024), 'Can gpt-4 models detect misleading visualizations?', *arXiv preprint arXiv:2408.12617* .

Aydın, Ö. and Karaarslan, E. (2023), 'Is chatgpt leading generative ai? what is beyond expectations?', *Academic Platform Journal of Engineering and Smart Systems* **11**(3), 118–134.

Azaria, A. (2023), Chatgpt: More human-like than computer-like, but not necessarily in a good way, *in* '2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)', IEEE, pp. 468–473.

Bashardoust, A., Feuerriegel, S. and Shrestha, Y. R. (2024), 'Comparing the willingness to share for human-generated vs. ai-generated fake news', *arXiv preprint arXiv:2402.07395* .

Bhattacharjee, A. and Liu, H. (2024), 'Fighting fire with fire: can chatgpt detect ai-generated text?', *ACM SIGKDD Explorations Newsletter* **25**(2), 14–21.

Bray, S. D., Johnson, S. D. and Kleinberg, B. (2023), 'Testing human ability to detect 'deepfake'images of human faces', *Journal of Cybersecurity* **9**(1), tyad011.

Cai, Z. G., Duan, X., Haslett, D. A., Wang, S. and Pickering, M. J. (2024), 'Do large language models resemble humans in language use?'.

Cao, S. and Huang, C.-M. (2022), 'Understanding user reliance on ai in assisted decision-making', *Proceedings of the ACM on Human-Computer Interaction* **6**(CSCW2), 1–23.

Castelo, N., Bos, M. W. and Lehmann, D. R. (2019), 'Task-dependent algorithm aversion', *Journal of Marketing Research* **56**(5), 809–825.

Chadha, A., Kumar, V., Kashyap, S. and Gupta, M. (2021), Deepfake: an overview, *in* 'Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020', Springer, pp. 557–566.

Chen, C. and Shu, K. (2023), 'Can llm-generated misinformation be detected?', *arXiv preprint arXiv:2309.13788* .

Chen, D. L., Schonger, M. and Wickens, C. (2016), 'otree—an open-source platform for laboratory, online, and field experiments', *Journal of Behavioral and Experimental Finance* **9**, 88–97.

Chiang, C.-H. and Lee, H.-y. (2023), 'Can large language models be an alternative to human evaluations?', *arXiv preprint arXiv:2305.01937* .

Chong, A. T. Y., Chua, H. N., Jasser, M. B. and Wong, R. T. (2023), Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features, *in* '2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)', IEEE, pp. 205–210.

Chu, S., Kim, J. and Yi, M. (2024), 'Think together and work better: Combining humans' and llms' think-aloud outcomes for effective text evaluation', *arXiv preprint arXiv:2409.07355* .

Desmond, M., Ashktorab, Z., Pan, Q., Dugan, C. and Johnson, J. M. (2024), Evalullm: Llm assisted evaluation of generative outputs, *in* 'Companion Proceedings of the 29th International Conference on Intelligent User Interfaces', pp. 30–32.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015), 'Algorithm aversion: people erroneously avoid algorithms after seeing them err.', *Journal of experimental psychology: General* **144**(1), 114.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2018), 'Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them', *Management science* **64**(3), 1155–1170.

Fagni, T., Falchi, F., Gambini, M., Martella, A. and Tesconi, M. (2021), 'Tweepfake: About detecting deepfake tweets', *Plos one* **16**(5), e0251415.

Friggeri, A., Adamic, L., Eckles, D. and Cheng, J. (2014), Rumor cascades, *in* 'proceedings of the international AAAI conference on web and social media', Vol. 8, No. 1, pp. 101–110.

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K. and Chen, L. (2023), 'Generative ai and chatgpt: Applications, challenges, and ai-human collaboration'.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S. and Wan, X. (2023), 'Human-like summarization evaluation with chatgpt', *arXiv preprint arXiv:2304.02554* .

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E. and Ghassemi, M. (2021), 'Do as ai say: susceptibility in deployment of clinical decision-aids', *NPJ digital medicine* **4**(1), 31.

Giglietto, F., Iannelli, L., Rossi, L. and Valeriani, A. (2016), 'Fakes, news and the election: A new taxonomy for the study of misleading information within the hybrid media system', *Giglietto, F., Iannelli, L., Valeriani, A., & Rossi, L.(2019).'Fake news' is the invention of a liar: How false information circulates within the hybrid news system. Current Sociology. DOI* **10**, 0011392119837536.

Gilovich, T., Medvec, V. H. and Savitsky, K. (2000), 'The spotlight effect in social judgment: an egocentric bias in estimates of the salience of one's own actions and appearance.', *Journal of personality and social psychology* **78**(2), 211.

Gino, F. (2008), 'Do we listen to advice just because we paid for it? the impact of advice cost on its use', *Organizational behavior and human decision processes* **107**(2), 234–245.

Greiner, B. (2015), 'Subject pool recruitment procedures: organizing experiments with orsee', *Journal of the Economic Science Association* **1**(1), 114–125.

Guo, Z. (2024), Online disinformation and generative language models: Motivations, challenges, and mitigations, *in* 'Companion Proceedings of the ACM on Web Conference 2024', pp. 1174–1177.

Harvey, N. and Fischer, I. (1997), 'Taking advice: Accepting help, improving judgment, and sharing responsibility', *Organizational behavior and human decision processes* **70**(2), 117–133.

Heckman, J. (1974), 'Shadow prices, market wages, and labor supply', *Econometrica: journal of the econometric society* pp. 679–694.

Heckman, J. E. and RajBhandary, U. L. (1979), 'Organization of trna and rrna genes in n. crassa mitochondria: intervening sequence in the large rrna gene and strand distribution of the rna genes', *Cell* **17**(3), 583–595.

Katarya, R. and Lal, A. (2020), A study on combating emerging threat of deepfake weaponization, *in* '2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)', IEEE, pp. 485–490.

Kaur, R. (2023), The impact of chatgpt on reliance on ai advice, Master's thesis, Universidade Catolica Portuguesa (Portugal).

Klingbeil, A., Grützner, C. and Schreck, P. (2024), 'Trust and reliance on ai—an experimental study on the extent and costs of overreliance on ai', *Computers in Human Behavior* **160**, 108352.

Koka, S., Vuong, A. and Kataria, A. (2024), 'Evaluating the efficacy of large language models in detecting fake news: A comparative analysis', *arXiv preprint arXiv:2406.06584* .

Kreps, S., McCain, R. M. and Brundage, M. (2022), 'All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation', *Journal of experimental political science* **9**(1), 104–117.

Kumar, S., Mahajan, S., Gadh, P., Jain, R. and Jain, D. (2024), 'Reliance of high school students on chat bots and its relationship with creativity', *Educational Research (IJMCER)* **6**(4), 286–290.

Kuosmanen, O. J. (2024), Advice from humans and artificial intelligence: Can we distinguish them, and is one better than the other?, Master's thesis, UiT Norges arktiske universitet.

LaGrandeur, K. (2021), 'How safe is our reliance on ai, and should we regulate it?', *AI and Ethics* **1**, 93–99.

Lee, J. and Shin, S. Y. (2022), 'Something that they never said: multimodal disinformation and source vividness in understanding the power of ai-enabled deepfake news', *Media Psychology* **25**(4), 531–546.

Leib, M., Köbis, N., Rilke, R. M., Hagens, M. and Irlenbusch, B. (2024), 'Corrupted by algorithms? how ai-generated and human-written advice shape (dis) honesty', *The Economic Journal* **134**(658), 766–784.

Lewandowsky, S., Ecker, U. K. and Cook, J. (2017), 'Beyond misinformation: Understanding and coping with the "post-truth" era', *Journal of applied research in memory and cognition* **6**(4), 353–369.

Li, Q., Cui, L., Kong, L. and Bi, W. (2023), 'Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation', *arXiv preprint arXiv:2310.19740* .

Logg, J. M., Minson, J. A. and Moore, D. A. (2019), 'Algorithm appreciation: People prefer algorithmic to human judgment', *Organizational Behavior and Human Decision Processes* **151**, 90–103.

Maggioni, M. A. and Rossignoli, D. (2023), 'If it looks like a human and speaks like a human... communication and cooperation in strategic human–robot interactions', *Journal of Behavioral and Experimental Economics* **104**, 102011.

Mai, K. T., Bray, S., Davies, T. and Griffin, L. D. (2023), 'Warning: Humans cannot reliably detect speech deepfakes', *Plos one* **18**(8), e0285333.

Mesbah, N., Tauchert, C. and Buxmann, P. (2021), 'Whose advice counts more–man or machine? an experimental investigation of ai-based advice utilization'.

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V. and Nguyen, C. M. (2022), 'Deep learning for deepfakes creation and detection: A survey', *Computer Vision and Image Understanding* **223**, 103525.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S. and Pollock, A. (2009), 'The relative influence of advice from human experts and statistical methods on forecast adjustments', *Journal of Behavioral Decision Making* **22**(4), 390–409.

Patil, M., Yadav, H., Gawali, M., Suryawanshi, J., Patil, J., Yeole, A. and Potlabattini, J. (2024), 'A novel approach to fake news detection using generative ai', *International Journal of Intelligent Systems and Applications in Engineering* **12**(4s), 343–354.

Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P. and Cherubini, A. (2022), 'Experimental evidence of effective human–ai collaboration in medical decision-making', *Scientific reports* **12**(1), 14952.

Sallami, D., Chang, Y.-C. and Aïmeur, E. (2024), 'From deception to detection: The dual roles of large language models in fake news', *arXiv preprint arXiv:2409.17416* .

Samarasinghe, K. and Prasangani, K. S. N. (2023), Reliance on ai tools and fostering creativity among sri lankan esl learners: Special focus to chatgpt, *in* '3rd International Conference on Educational Technology and Online Learning', pp. 97–101.

Sareen, M. (2022), Threats and challenges by deepfake technology, *in* 'DeepFakes', CRC Press, pp. 99–113.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C. and Satzger, G. (2022), 'Should i follow ai-based advice? measuring appropriate reliance in human-ai decision-making', *arXiv preprint arXiv:2204.06916* .

Schemmer, M., Kuehl, N., Benz, C., Bartos, A. and Satzger, G. (2023), Appropriate reliance on ai advice: Conceptualization and the effect of explanations, *in* 'Proceedings of the 28th International Conference on Intelligent User Interfaces', pp. 410–422.

Schuff, H., Vanderlyn, L., Adel, H. and Vu, N. T. (2023), 'How to do human evaluation: A brief introduction to user studies in nlp', *Natural Language Engineering* **29**(5), 1199–1222.

Shekar, S., Pataranutaporn, P., Sarabu, C., Cecchi, G. A. and Maes, P. (2024), 'People over trust ai-generated medical responses and view them to be as valid as doctors, despite low accuracy', *arXiv preprint arXiv:2408.15266* .

Sniezek, J. and Buckley, T. (1989), Social influence in the advisor-judge relationship, *in* 'Annual meeting of the judgment and decision making society, Atlanta, Georgia'.

Takale, D. G., Mahalle, P. N. and Sule, B. (2024), 'Cyber security challenges in generative ai technology', *Journal of Network Security Computer Networks* **10**(1), 28–34.

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S. et al. (2024), 'A framework for human evaluation of large language models in healthcare derived from literature review', *NPJ Digital Medicine* **7**(1), 258.

Tinghu, K., Li, W., Peiling, X. and Qian, P. (2018), 'Taking advice for vocational decisions: Regulatory fit effects', *Journal of Pacific Rim Psychology* **12**, e7.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A. and Ortega-Garcia, J. (2020), 'Deepfakes and beyond: A survey of face manipulation and fake detection', *Information Fusion* **64**, 131–148.

Tse, T. T. K., Hanaki, N. and Mao, B. (2024), 'Beware the performance of an algorithm before relying on it: Evidence from a stock price forecasting experiment', *Journal of Economic Psychology* **102**, 102727.

Tseng, R., Verberne, S. and van der Putten, P. (2023), Chatgpt as a commenter to the news: can llms generate human-like opinions?, *in* 'Multidisciplinary International Symposium on Disinformation in Open Online Media', Springer, pp. 160–174.

Uchendu, A., Lee, J., Shen, H., Le, T., Lee, D. et al. (2023), Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts?, *in* 'Proceedings of the AAAI Conference on Human Computation and Crowdsourcing', Vol. 11, No. 1, pp. 163–174.

Uchendu, A., Venkatraman, S., Le, T. and Lee, D. (2024), Catch me if you gpt: Tutorial on deepfake texts, *in* 'Proceedings of the 2024 Conference of the North American

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)', pp. 1–7.

Vodrahalli, K., Daneshjou, R., Gerstenberg, T. and Zou, J. (2022), Do humans trust advice more if it comes from ai? an analysis of human-ai interactions, *in* 'Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society', pp. 763–777.

Waisbord, S. (2018), 'Truth is what happens to news: On journalism, fake news, and post-truth', *Journalism studies* **19**(13), 1866–1878.

Walkowiak, E. and MacDonald, T. (2023), 'Generative ai and the workforce: What are the risks?', *Available at SSRN* .

Wang, M. (2024), 'Generative ai: A new challenge for cybersecurity', *Journal of Computer Science and Technology Studies* **6**(2), 13–18.

Whyte, C. (2020), 'Deepfake news: Ai-enabled disinformation as a multi-level public policy challenge', *Journal of cyber policy* **5**(2), 199–217.

Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M. and Choi, Y. (2021), 'Reframing human-ai collaboration for generating free-text explanations', *arXiv preprint arXiv:2112.08674* .

Yaniv, I. (2004), 'The benefit of additional opinions', *Current directions in psychological science* **13**(2), 75–78.

Yaniv, I. and Foster, D. P. (1997), 'Precision and accuracy of judgmental estimation', *Journal of behavioral decision making* **10**(1), 21–32.

Yoon, H., Scopelliti, I. and Morewedge, C. K. (2021), 'Decision making can be improved through observational learning', *Organizational Behavior and Human Decision Processes* **162**, 155–188.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F. and Choi, Y. (2019), 'Grover-a state-of-the-art defense against neural fake news', *Proc. Adv. Neural Inf. Process. Syst* **32**.

Zhang, C., Zhang, C., Li, C., Qiao, Y., Zheng, S., Dam, S. K., Zhang, M., Kim, J. U., Kim, S. T., Choi, J. et al. (2023), 'One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era', *arXiv preprint arXiv:2304.06488* .

Zhu, T., Weissburg, I., Zhang, K. and Wang, W. Y. (2024), 'Human bias in the face of ai: The role of human judgement in ai generated text evaluation', *arXiv preprint arXiv:2410.03723* .

A Quiz Questions

We set four true or false questions to make sure participants basically understand the rules and the answers are $true$, $true$, $false$, $false$.

Q1: In each news, the minimum value of the fraction of the real news part is 0, and the maximum value is 100.

Q2: In this experiment, the 'authenticity' you are asked to identify can actually be considered as 'the fraction of the real news part'.

Q3: Additional payoff besides the participation fee are related to the accuracy of your identifications, and the total additional reward is the sum of the rewards for all your identifications.

Q4: After making the first identification and receiving the advice of ChatGPT's identification, you must make a second identification different from the first one. (if in **AI group**)

Q4: After making the first identification and receiving the advice of someone else's first identification, you must make a second identification different from the first one. (if in **Human group**)

B Experiment Screen

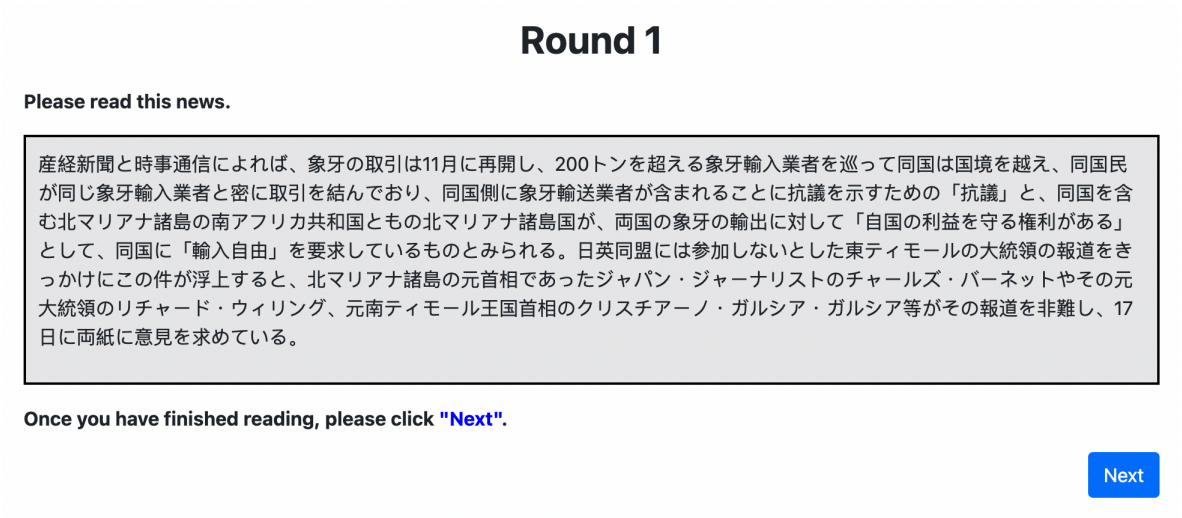

Figure 11: Read News

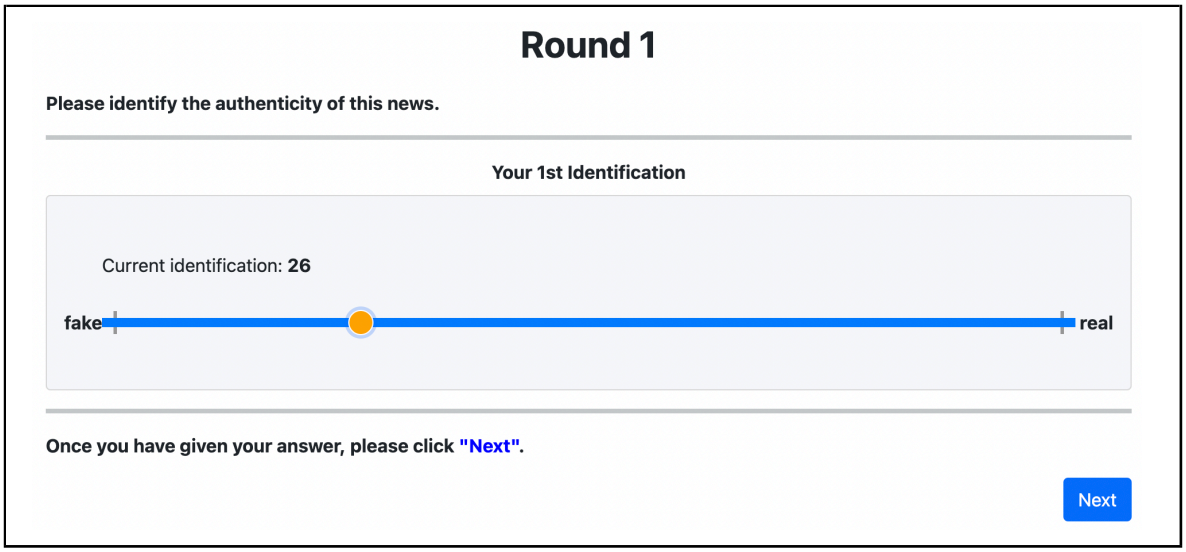

Figure 12: First Identification

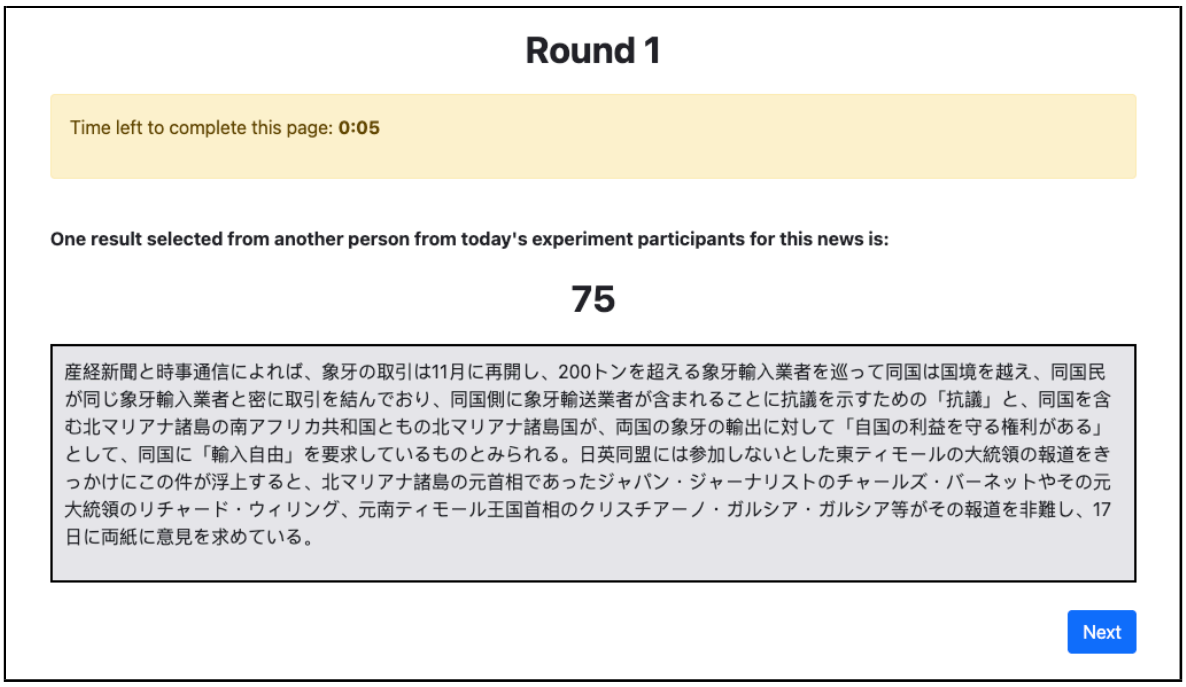

Figure 13: Human Peer's advice

The translation of the news text in Figures 9 and 11 is as follows:

According to Sankei Shimbun and Jiji Press, the ivory trade is set to resume in November, involving over 200 tons of ivory. This issue has led to cross-border disputes, with nationals of a particular country closely engaging with the same ivory importers and including ivory transporters from that country, prompting protests. Additionally, countries including the Northern Mariana Islands and the Republic of South Africa argue that they have the right to protect their interests regarding ivory exports and are demanding "free import" from the concerned country. The issue surfaced following reports that the President of East Timor decided not to join the Anglo-Japanese Alliance. Former Prime Minister of the Northern Mariana Islands, Japan Journalist Charles Barnett, former President Richard Willing, and former Prime Minister of the Kingdom of South Timor, Cristiano Garcia Garcia, have criticized these reports. Both newspapers have been asked for their opinions on the 17th.

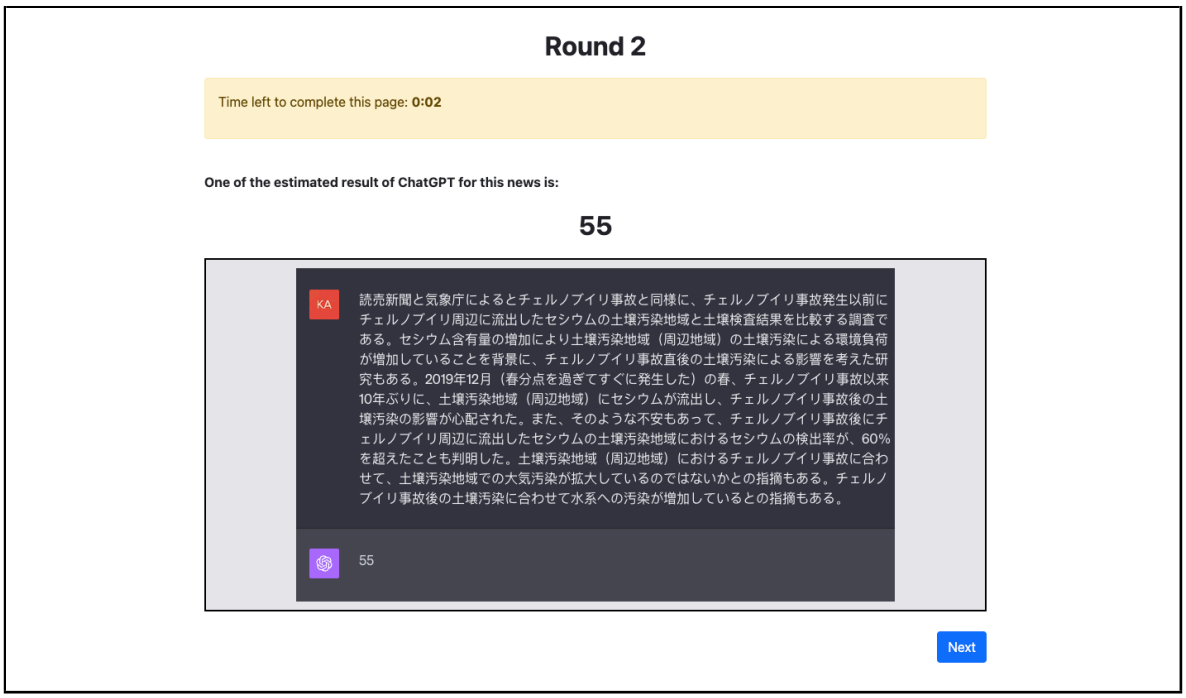

Figure 14: ChatGPT's advice

The translation of the news text in Figure 12 is as follows:

According to Yomiuri Shimbun and the Meteorological Agency, the study compares soil contamination with cesium in areas around Chernobyl before the Chernobyl accident to the results of soil tests conducted after the accident. The increase in cesium content has intensified the environmental burden due to soil contamination in the surrounding areas. There are also studies considering the impact of soil contamination immediately following the Chernobyl accident. In the spring of 2019, shortly after the vernal equinox, cesium leaked into the soil-contaminated areas around Chernobyl for the first time in ten years since the accident, raising concerns about the effects of post-accident soil contamination. Due to such concerns, it was also discovered that the detection rate of cesium in the soil-contaminated areas around Chernobyl exceeded 60%. There are indications that air pollution in the soil-contaminated areas may be expanding in line with the Chernobyl accident. Additionally, there are concerns that water pollution has been increasing in line with the soil contamination after the Chernobyl accident.

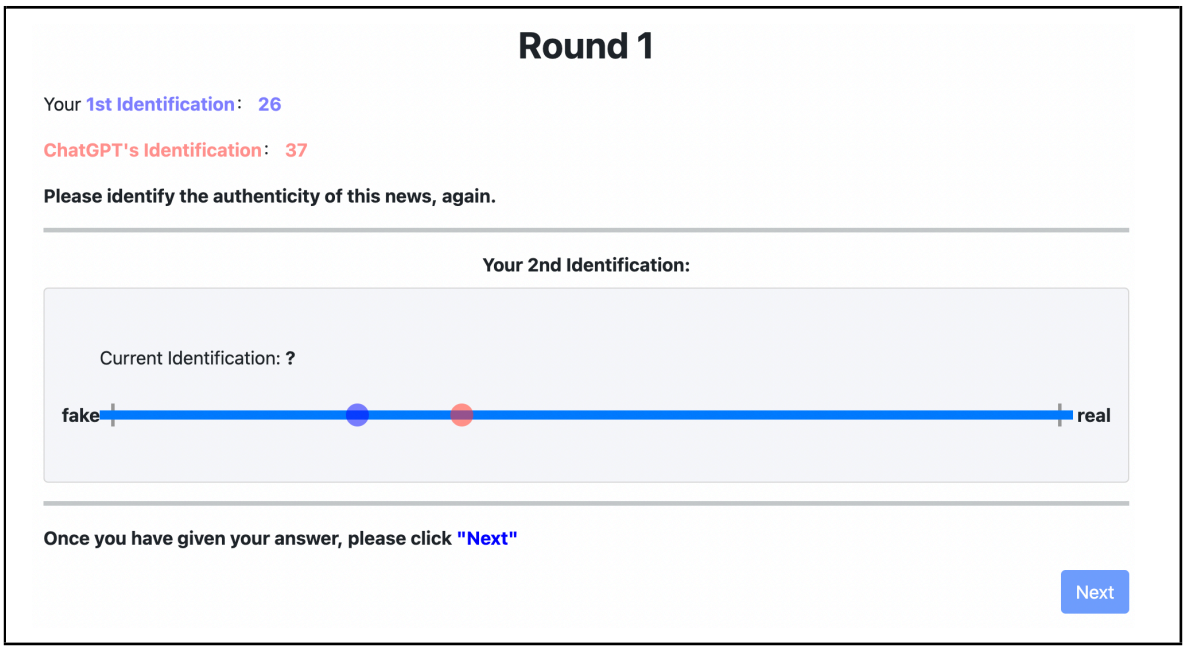

Figure 15: Second Identification

C Survey Questions

The survey questions used in our experiment are as follows.

SQ1: Please input your age: [ ]

SQ2: Please select your gender: [male/female/other/not want to answer]

SQ3: Which college or research institute are you affiliated with?

SQ4: After making the first identification and receiving the advice of ChatGPT's identification, you must make a second identification different from the first one.

SQ5: Have you heard about ChatGPT?

SQ6: How many days per week do you use ChatGPT on average?

SQ7: In today's experiment, specifically in the task of "assessing News' authenticity," who do you think can provide more accurate responses?

D Examples of Utilization Statuses

The following plots show examples of the three utilization statuses described in Section 5.3. Each example displays a slider used in the second identification stage, where a participant's first identification (blue point), second identification (orange point), and advice (red point) for that round are marked on the slider.

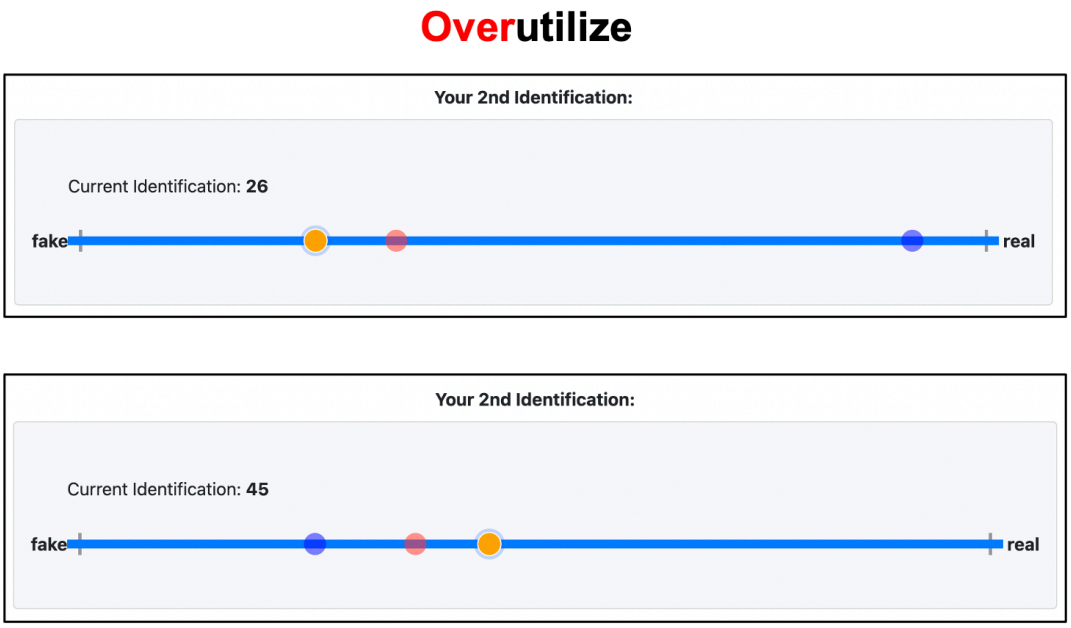

Figure 16: Two Examples of Over-utilize

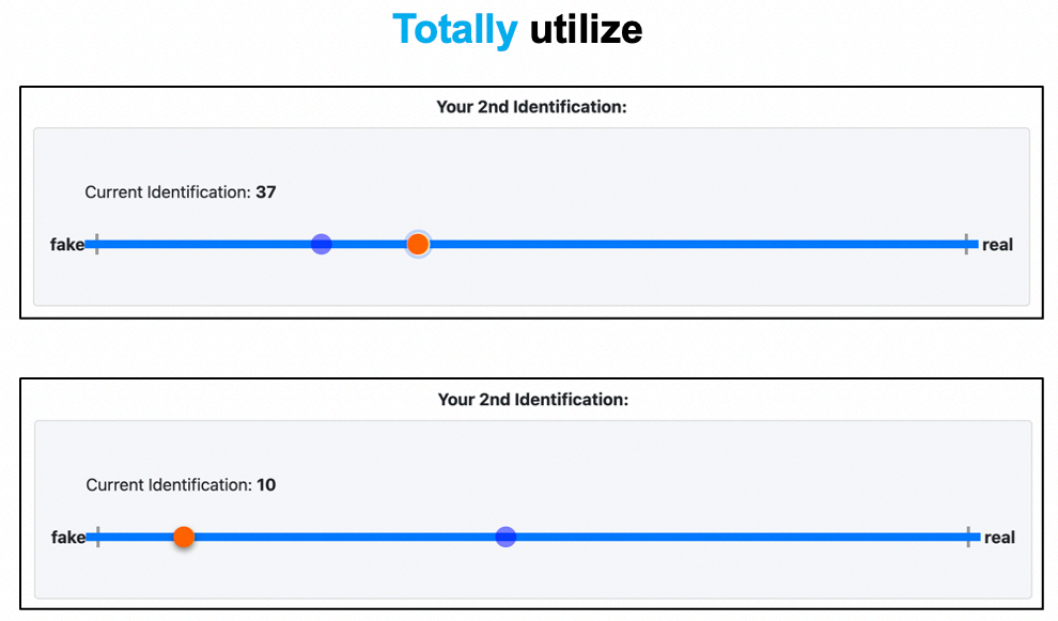

Figure 17: Two Examples of Totally Utilize

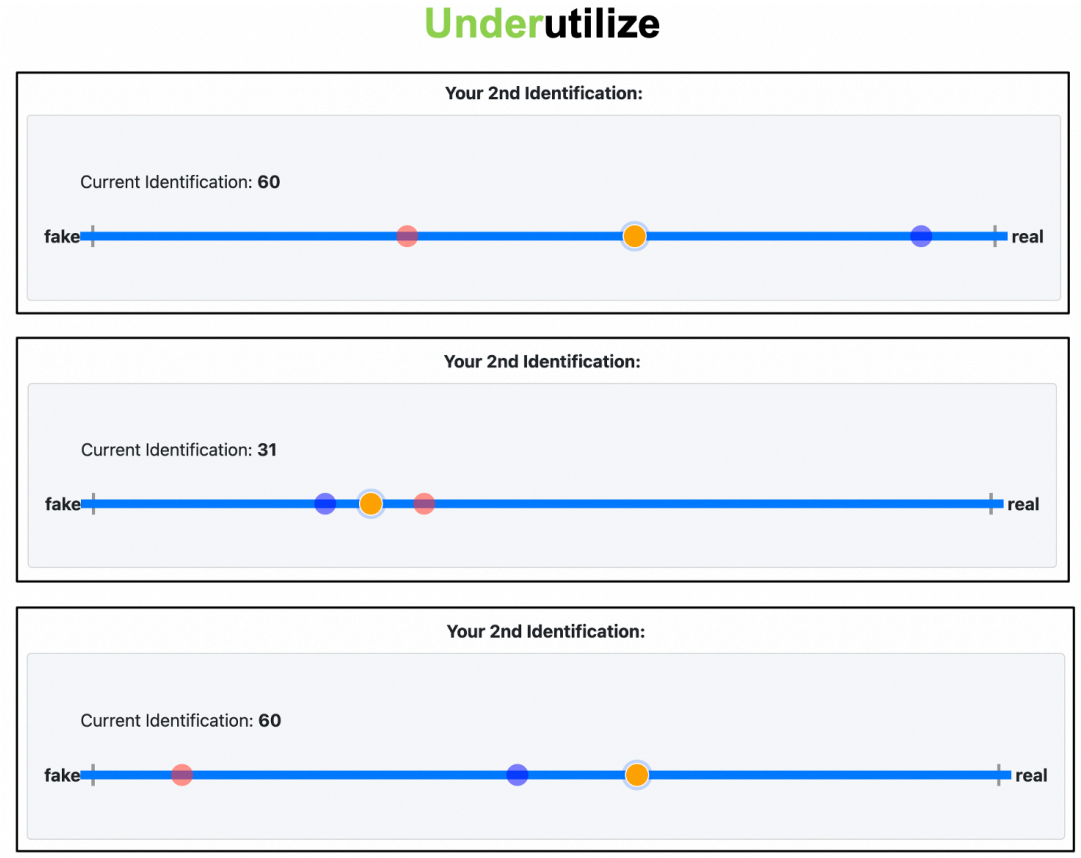

Figure 18: Three Examples of Under-utilize