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

Secondly Revised February 2026
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[‡]

February 5, 2026

Abstract

This experimental study investigates how people rely on different sources of advice when detecting AI-generated fake news (deepfake news). In a laboratory deepfake detection task, student participants identified the proportion of human-written (non-AI-generated) content in synthetic deepfake news articles and received advice from ChatGPT (GPT-4), human peers, or linguistic experts. The results show that participants rely more on ChatGPT than on human peers when detecting GPT-2-generated deepfake news. Participants also rely more on linguistic experts than on peers, while the relative reliance on experts versus ChatGPT is mixed across experimental waves, potentially reflecting time trends in beliefs about AI-based detection. Moreover, performance improvements reflect the joint role of reliance and advice quality, arising primarily when participants rely on high-quality advice. Overall, relying on AI to detect AI-generated deepfakes can improve detection outcomes, but only when AI-based detection tools are of sufficiently high quality. These findings highlight the dual role of GAI as both a source of deepfakes and a tool for mitigating related risks.

Keywords: ChatGPT, AI reliance, deepfake detection, advice taking, human–AI interaction

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 the University of Osaka; (b) Grants-in-Aid for Scientific Research (Nos. 20H05631, 23H00055, and 25H00388) KAKENHI from the Japan Society for the Promotion of Science; and (c) the Support for Pioneering Research Initiated by the Next Generation program of the Japan Science and Technology Agency (No. JPMJSP2138). 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 thank Tiffany Tsz Kwan Tse and Taisuke Imai for their helpful comments and suggestions. We are also grateful to Bor Hodošček, Takahiro Honda, Hiroshi Mito, Masao Ochi, Yushi Sugimoto, Eri Tanaka, and Yoko Yumoto for providing advice materials in the additional experiment, and to those experts who wished to remain anonymous. Finally, we gratefully acknowledge the support of Yuta Shimodaira, Zhelin Zhang, and Satsuki Yamada in conducting the experiment.

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

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

1 Introduction

Generative AI (GAI) products, such as ChatGPT, have attracted widespread attention since 2022 and have been rapidly adopted across many domains. Alongside this diffusion, growing concerns have emerged regarding their societal harms, particularly the spread of AI-generated misinformation (deepfakes) (Lundberg and Mozelius, 2024; Sophia, 2025). In response, recent studies and policy discussions increasingly propose leveraging GAI itself—such as commercial or platform-integrated AI detectors (e.g., Turnitin)—to identify and mitigate deepfake content (Patil et al., 2024; Bhattacharjee and Liu, 2024). However, these tools are imperfect and can misclassify human-written text as AI-generated, producing false positives with tangible real-world consequences (Weber-Wulff et al., 2023; Knilans, 2024; Zhang et al., 2024a). GAI thus increasingly plays a dual role: *it is both a source of deepfakes and an imperfect tool used to combat them.*

This dual role highlights the importance of understanding individuals' reliance on GAI—based tools to detect deepfake content—whether they choose to "fight fire (AI) with fire (AI)," and whether they appropriately account for both the benefits and risks of doing so. At the same time, such reliance may itself become problematic, as concerns about over-reliance on AI systems have grown substantially (Schemmer et al., 2023; Klingbeil et al., 2024).

Building on this context, this study examines individuals' "reliance on AI to detect AI". Specifically, we address the following research question:

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

We conducted a laboratory experiment in which participants performed a deepfake detection task. Specifically, they were asked to identify the proportion of non–AI-generated (human-written) content in synthetic deepfake news articles composed of both human-written and AI-generated text. After providing an initial identification, participants in our main experiments received advice in the form of corresponding identifications generated either by ChatGPT (GPT-4) or by human peers who participated in the same experimental session.

We adopt the "weight of advice" (WOA) measure (Harvey and Fischer, 1997; Yaniv and Foster, 1997) to quantify the reliance and find that participants rely more on ChatGPT than on human peers when detecting deepfake news. Moreover, advice from ChatGPT leads to higher performance, reflecting its higher advice quality. Importantly, variation in the proportion of human-written (non-AI-generated) content within articles does not systematically affect either reliance behavior or performance.

We also conducted an additional experiment that introduced two new advice sources—*linguistic experts* and *human peers from earlier experimental sessions*—while also replicating the condition in which participants received advice from ChatGPT but varying the timing of the questionnaire measuring participants' prior beliefs, moving it from after the main task to before the task, to test whether questionnaire placement affects reported beliefs. The results show that questionnaire timing does not significantly affect prior beliefs, and that participants' reliance on human peers does not differ depending on whether peers participated on the same day or in earlier sessions. Participants rely more on linguistic experts than on human peers, while the relative reliance on experts compared to AI is mixed—lower than reliance on AI in the main experiment but higher in the additional experiment. We interpret this pattern as reflecting time trends in beliefs about AI-based deepfake detection.

In addition, we examine how participants' prior beliefs—capturing their subjective preferences over advice sources—and advice quality jointly shape performance improvement. Our results indicate that improvements in detection accuracy arise from the interaction between advice quality and reliance: participants rely more on sources they personally trust, and the benefits of high-quality advice materialize only when advice is actually used.

Because the WOA measure does not totally utilize all observations—and also as a robustness check for the main analysis—we further adopt the two-stage framework of Vodrahalli et al. (2022) and implement it using a Heckman selection correction model (Heckman, 1974, 1979). This approach decomposes reliance into an "activation" stage (whether advice is taken) and an "integration" stage (how strongly it is incorporated). The results validate this decomposition: advice source and prior beliefs systematically shape both stages, while the advice–initial gap (gap between the advice and initial response) exerts opposite effects—encouraging activation but dampening integration.

These findings have direct implications for policy frameworks aimed at mitigating the risks of deepfake misinformation. They suggest that, while people tend to trust AI-based detection tools, the effective use of such tools depends not only on public beliefs and trust but also on their objective detection quality. Accordingly, regulators and institutions may consider promoting the development and responsible deployment of AI-based detection systems, while recognizing that policy outcomes are jointly shaped by detectors' performance and users' beliefs. In this sense, encouraging the adoption of high-quality AI-based detection tools may be a more effective policy priority than relying on source credibility or trust alone, without regard to realized detection performance.

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 results of the main experiments. **Section** 5 reports the additional experiments. **Section** 6 provides the discussions of the findings, and **Section** 7 concludes the paper.

2 Literature Review

We review two strands of research. First, we summarize the literature on deepfake detection, focusing on its societal relevance and detection challenges. Second, we review prior work on reliance on AI advice, with attention to behavioral mechanisms, measurement approaches, and applications across domains.

2.1 Background on Deepfake Detection

2.1.1 Why Deepfakes Matter

Deepfakes refer to synthetic media—primarily audio and video—generated using deep learning techniques (Chadha et al., 2021). These outputs are designed to closely mimic real content, which makes them potentially harmful to society (Katarya and Lal, 2020; Sareen, 2022). With the rapid advancement of GAI, the concept of deepfakes has expanded beyond audio and video to include textual content (Chong et al., 2023; Uchendu et al., 2023, 2024), making the creation and spread of deepfakes increasingly easy and widespread.

Deepfakes pose serious risks across multiple domains. In politics, they can increase cognitive load, reinforce confirmation bias, and undermine social trust, posing particular threats to elections and democratic processes (Islam et al., 2024; Amin et al., 2025; Gupta et al., 2025). In academia and education, the use of deepfake data and images creates systemic risks to research integrity (Chen et al., 2024b; Chauhan and Currie, 2024). Deepfakes also threaten individuals' daily lives by enabling fraud, identity misuse, and reputational harm, such as deepfake pornography created using others' images (Umbach et al., 2024) and voice-based scams using synthetic speech (Barrington et al., 2025).

In response, scholars emphasize the need for coordinated efforts across policy, technology, and human decision-making: governments are urged to develop targeted legal responses (Yamaoka-Enkerlin, 2019; Ramluckan, 2024), researchers continue to improve detection methods and tools (Mirsky and Lee, 2021), and users increasingly seek ways to verify media in everyday social contexts as deepfakes become harder to distinguish from real content (Ahmed and Chua, 2023). Complementing detection-based approaches, recent work also explores human-centered interventions that aim to mitigate deepfake harms by shaping beliefs and source perceptions prior to exposure, such as pre-emptive source discreditation (i.e., warning users about the unreliability of a source before exposure) and debunking of AI-generated misinformation (Spearing et al., 2025).

2.1.2 Human and Machine Detection of Deepfakes

Growing evidence shows that humans perform poorly when trying to detect deepfake content. Chen and Shu (2023) find that AI-generated misinformation—such as outputs from GPT-3.5—is harder for people to identify than human-written fake news, making it potentially more harmful. A systematic review and meta-analysis by Diel et al. (2024) reaches a similar conclusion: across 56 studies, average detection accuracy is only 55.54% (and just 52% for deepfake text). Consistent with these findings, Groh et al. (2024) report that increasing the share of deepfakes in a set of political speeches does not meaningfully change people's judgments, and that deepfake text is even harder to detect than manipulated audio or video.

Compared with human detection, much of the recent literature focuses on machine-based approaches to deepfake detection. As Zellers et al. (2019) argue, "the best way to detect neural fake news is to use a model that is also a generator." Empirical studies provide partial support for this view. For example, Alexander et al. (2024) show that GPT-4 can identify misleading visualizations with moderate accuracy even without task-specific training, and Koka et al. (2024) report detection accuracy exceeding 95% for GPT-4 on deepfake news. However, detection performance is far from uniform. Sallami et al. (2024) find that while GPT-4 performs well on AI-generated misinformation, it is substantially less accurate when detecting human-created fake news. Public AI detectors exhibit even greater instability: although some benchmark studies report high accuracy rates (Koka et al., 2024; Sallami et al., 2024; Liu et al., 2024), others document performance only slightly above chance. In particular, Weber-Wulff et al. (2023) show that widely used detectors—including commercial systems such as Turnitin—are easily fooled by paraphrasing, a conclusion echoed by the review in Chaka (2024). As GAI continue to advance, both deepfake generation and detection technologies evolve rapidly, creating an ongoing "generation–detection" arms race (Laurier et al., 2024).

Importantly, these performance limitations are not merely technical but can translate into real-world risks. A growing body of evidence documents that AI detectors frequently misclassify human-written text as AI-generated, leading to false positives (Weber-Wulff et al., 2023; Knilans, 2024; Zhang et al., 2024a). Even OpenAI cautions that AI-writing detectors are not reliable for high-stakes judgments.[1] Well-known cases include detectors labeling the *U.S. Constitution* and passages from the *Bible* as AI-generated,[2] and false positives have led to tangible harms in education, such as publicized misconduct cases and delayed graduations (Bergin, 2025; Hallikaar, 2025; The Advertiser, 2025; The Courier-Mail, 2024).

Beyond individual performance, collaboration has been shown to substantially enhance human deepfake detection. *Human–Human collaboration* improves accuracy: Uchendu et al. (2023) find that group discussion leads to better detection of deepfake text than individual judgment for both experts and non-experts, and Groh et al. (2022) show that aggregating human predictions on deepfake videos yields accuracy comparable to state-of-the-art detectors and clearly above that of single raters. These effects align with the review in Diel et al. (2024). *Collaboration with AI* appears similarly promising: Experiments in Groh et al. (2022) show that participants who observe an

[1] OpenAI Help Center: "Do AI detectors work? In short, not in our experience."

[2] See Ars Technica on the Constitution case (Edwards, 2023) and India Today on Bible passages being flagged as AI (Chakravarti, 2023).

AI model's prediction outperform both standalone humans and the model itself. Diel et al. (2024) likewise note systematic accuracy gains when humans receive AI assistance, and Somoray et al. (2025) demonstrate that humans and models rely on different cues when judging authenticity—suggesting complementarities that human–AI collaboration can leverage.

2.1.3 Experimental Innovations

While deepfakes pose serious societal risks and collaboration with AI or other humans can improve detection accuracy, neither human nor machine detection is perfect. As a result, uncritical reliance on external detection results may introduce secondary risks, particularly in high-stakes contexts. This tension motivates our examination of how individuals rely on AI tools versus human collaboration when detecting deepfake news.

In this study, we focus on a common form of deepfake—deepfake news—and examine a deepfake detection task in a laboratory setting in which participants identify the proportion of non-AI-generated text in each article. Although this design cannot fully replicate real-world environments, it provides an initial and controlled way to measure how individuals perceive deepfake content and how they behave when detecting it.

Importantly, the task is designed to capture detection of AI-generated content embedded in deepfake news articles, rather than the identification of human-written fake news per se. Most experimental studies in economics on human-written fake news detection ask participants for a binary judgment (real vs. fake) (Serra-Garcia and Gneezy, 2021; Arin et al., 2023; Thaler, 2024). By contrast, we ask participants to report the article's non-AI-generated proportion (human-written proportion), following recent work that moves beyond whole-document labels toward partial detection and localization (Zeng et al., 2024; Zhang et al., 2024b). We use a proportion for four reasons. First, our stimuli include both totally AI-generated and totally human-written items, so a proportion nests binary judgments while offering finer measurement. Second, as models improve, binary human detection becomes unreliable, whereas a proportion better captures subjective uncertainty. Third, real-world content often mixes AI output with human editing, making proportion ratings more aligned with actual production processes. Fourth, modern AI detectors themselves output continuous scores or estimated "AI-generated proportions," rather than hard labels.

We incorporate both human–human and human–AI collaboration. Participants receive advice either from peers' initial identifications or from identifications generated by ChatGPT. This design enables us, for the first time in the deepfake-detection context, to directly compare how individuals rely on and respond to these two distinct modes of collaboration.

We also introduce a third source of advice: linguistic experts. Because the development of GAI—especially large language models (LLMs)—relies heavily on linguistic knowledge and analysis, linguists and linguistically trained annotators are frequently involved in dataset curation, model evaluation, and assessments of whether LLM-generated text resembles human-written language (Bender et al., 2021; Ouyang et al., 2022; Workshop et al., 2022). Introducing this additional advice source allows us to investigate a previously unexplored question: *do people trust experts who are familiar with, and in some cases involved in, the development of LLMs to accurately identify AI-generated text*? This aspect of human judgment has not been examined in existing deepfake detection or experimental economic studies.

2.2 Previous Work on AI Reliance

2.2.1 Algorithm Aversion, Appreciation, and Human Perception of AI

A large body of research examines how individuals respond to algorithmic advice relative to human judgment. A central finding is that people do not treat algorithmic and human advice symmetrically: reliance on algorithms varies systematically across contexts, ranging from algorithm aversion, where individuals rely less on algorithms than on humans (Dietvorst et al., 2015), to algorithm appreciation, where algorithmic judgments receive greater weight (Logg et al., 2019).

Evidence of "algorithm aversion" has been documented across a range of settings. For example, experts often rely less on algorithmic systems than on non-expert humans (Reverberi et al., 2022; Agarwal et al., 2023), and similar patterns arise when algorithms are compared with human peers rather than experts (Gaube et al., 2021; Mesbah et al., 2021). In contrast, Logg et al. (2019) show that individuals may rely more on algorithms than on humans when algorithms are perceived as appropriate for the task. Related work also finds that allowing users to make small adjustments to algorithmic advice can reduce algorithm aversion and increase reliance (Dietvorst et al., 2018). These mixed findings indicate that algorithm aversion is not a fixed bias but depends on how algorithms are perceived and used.

Subsequent research explores the mechanisms underlying these divergent responses. Castelo et al. (2019) show that reliance on algorithms depends on task characteristics, particularly perceived objectivity: algorithm aversion is stronger in subjective tasks than in objective ones. They further find that increasing the perceived human likeness of algorithms can mitigate aversion in subjective domains. Other studies emphasize the role of interaction and control. For instance, Maggioni and Rossignoli (2023) show that verbal interaction with robots reduces algorithm aversion, while Tse et al. (2024) find that granting decision makers greater freedom in final decisions can induce over-reliance on algorithms, even when algorithmic performance is low. Finally, focusing on process design in high-stakes settings, Yin et al. (2025) examine how the timing of AI advice affects diagnostic decision making and find that physicians perform best when AI advice is provided after an initial diagnosis, and worst when no AI advice is available.

Recent work further synthesizes these findings by focusing on how individuals conceptualize and evaluate AI itself. As AI technologies have evolved, the notion of AI has expanded beyond task-specific algorithms to encompass a broad set of interrelated technologies, including algorithms, decision-support systems, social robots, and conversational agents such as ChatGPT (Walsh et al., 2019; Baines et al., 2024). In particular, Baines et al. (2024) review a wide literature on advice from AI across domains, emphasizing how trust, acceptance, and reliance on AI advice depend on contextual and individual factors. Complementing this perspective, Passi et al. (2025) synthesize evidence from over 120 interdisciplinary studies to examine the negative consequences of AI mistakes, with particular attention to overreliance on AI—especially GAI—and the antecedents and mitigation strategies associated with such overreliance. Similarly, Chevrier (2025) propose a conceptual framework that organizes prior evidence around key dimensions such as perceived competence, accountability, and controllability in the advice-taking process, especially for ChatGPT. Consequently, this integrative body of work suggests that insights from earlier studies on human-algorithm interaction can be naturally extended to contemporary research on human–AI interaction, including reliance on GAI systems.

2.2.2 Applications of AI Reliance Across Domains

More recently, research on reliance on AI advice has expanded to a broader set of application domains. In economic and decision-making settings, recent work examines how individuals rely on AI advice across diverse contexts, including strategic games (Klingbeil et al., 2024), financial forecasting and investment decisions (Kim and Park, 2024), organizational and managerial decision making (Stiefenhofer et al., 2026), and labor-related tasks shaped by GAI (Brynjolfsson et al., 2025).

Beyond economics, medical contexts constitute a major area of study on reliance on AI advice across multiple stages of clinical decision making. Existing work examines how AI advice affects diagnostic processes and information integration (Yin et al., 2025), as well as trust, acceptance, and explainability of medical AI systems (Mainz, 2024; Küper et al., 2025; Rosenbacke et al., 2024; Tun et al., 2025). A growing literature also focuses on patients' perceptions of AI- and ChatGPT-based healthcare advice, documenting how trust and acceptance vary across tasks and informational settings (Chen et al., 2024a; Sun et al., 2024; Van Bulck and Moons, 2024; Chen and Cui, 2025; Kelly et al., 2025).

Emerging evidence from other domains, including law (Tamò-Larrieux et al., 2024) and education (Viberg et al., 2025; Amoozadeh et al., 2024), further suggests that reliance on AI advice is a domain-general phenomenon shaped by institutional context and task structure.

2.2.3 Measuring AI Reliance

The judge–advisor paradigm (JAS) (Sniezek and Buckley, 1989) is widely used in economics and psychology to study the reliance on algorithm. In this framework, the reliance is commonly quantified using the WOA, which measures the extent to which individuals adjust their initial judgments toward the advice. A comprehensive meta-analysis by Bailey et al. (2023) shows that individuals, on average, place substantially less than equal weight on advice and that WOA varies systematically with task characteristics, advisor attributes, and decision makers' confidence, underscoring that WOA captures a behavioral response rather than a stable preference parameter. In studies of algorithm aversion, WOA has been widely used to quantify reliance on algorithmic advice (Dietvorst et al., 2015, 2018; Logg et al., 2019), and more recent work extends this approach to measure reliance on GAI, including LLMs such as ChatGPT (Zhang, 2023; Rebholz et al., 2024; Boob-Engel, 2025; Bo et al., 2025).

Beyond behavioral advice-taking measures, a large literature studies responses to AI advice using survey-based approaches. Many studies adopt acceptance-oriented frameworks such as the Technology Acceptance Model (TAM), which focus on perceived usefulness, ease of use, and intentions to adopt AI systems (Davis, 1989; Venkatesh et al., 2003). For example, Biswas and Murray (2024) examine how educational background and self-perceived technological proficiency relate to reliance on AI, and Setyaningsih et al. (2025) study students' reliance on ChatGPT's writing suggestions within a TAM-based framework. Other survey-based studies applied direct elicitation, asking individuals to report whether, or to what extent, they rely on AI advice, sometimes using vignette-based designs (Hoff and Bashir, 2015; Bussone et al., 2015; Glikson and Woolley, 2020; Rosenbacke et al., 2024). Such approaches are widely used in applied domains, including medicine (Bussone et al., 2015), law (Eckhardt et al., 2025), and management (Glikson and Woolley, 2020). While informative about attitudes, intentions, and stated use of AI advice, these survey-based measures do not directly capture how advice is integrated into incentivized decision making.

2.2.4 Benchmarking AI Advice: Human Peers and Experts

Across these approaches, reliance on AI advice is typically evaluated relative to benchmark sources of judgment, most commonly human peers or domain experts. These benchmarks serve different purposes and correspond to distinct research questions.

Human peers provide a natural and methodologically conservative benchmark for evaluating reliance on AI advice. Because peer judgments reflect lay people's decision making, they allow researchers to isolate the effect of advice source without conflating it with differences in expertise or authority. Using this benchmark, prior studies compare reliance on AI advice with reliance on advice from non-expert humans holding objective quality constant. For example, Vodrahalli et al. (2022) show that individuals are more likely to activate advice labeled as AI than otherwise identical advice attributed to human peers. Similarly, Zheng et al. (2025) compare reliance on AI advice and human peer advice in numerical estimation tasks and find no systematic tendency toward greater reliance on AI than on peers. Despite these contributions, direct comparisons between human peers and GAI systems—particularly ChatGPT—remain relatively limited, leaving open questions about how reliance on GAI differs from reliance on ordinary human judgment.

In contrast, domain experts provide a normative benchmark, especially in high-stakes or specialized tasks where decision quality rather than conformity to typical human judgment is the primary concern. A growing literature compares reliance on AI advice with reliance on expert advice across domains. For instance, Agrawal et al. (2023) examine reliance on AI versus human experts across OECD and Indian samples, finding that participants evaluate experts and AI differently along dimensions of competence, trust, and moral responsibility. Larkin et al. (2022) study reliance on AI and human expert advice in medicine and finance and show that participants update their decisions more in response to expert recommendations, although the magnitude of this effect varies by context. However, direct evidence comparing reliance on GAI systems such as ChatGPT with reliance on human experts remains scarce, particularly in controlled experimental settings that allow for clean measurement of advice integration.

2.2.5 Contributions

Relative to the existing literature on AI reliance, this study makes three main contributions.

First, we study reliance on AI advice in the context of deepfake detection, a setting that has become increasingly salient with the rapid improvement of GAI technologies. Our design allows us to examine how individuals rely on AI assistance when evaluating content generated by AI itself, highlighting the dual role of GAI. In light of the rapid spread of deepfakes and the growing use of AI-based detection tools in real-world contexts, our analysis helps inform more effective approaches to the governance and regulation of deepfake content.

Second, we directly compare advice from ChatGPT with advice from human peers and domain experts within a unified experimental framework. Although recent studies examine reliance on AI advice in isolation or relative to a single benchmark, systematic

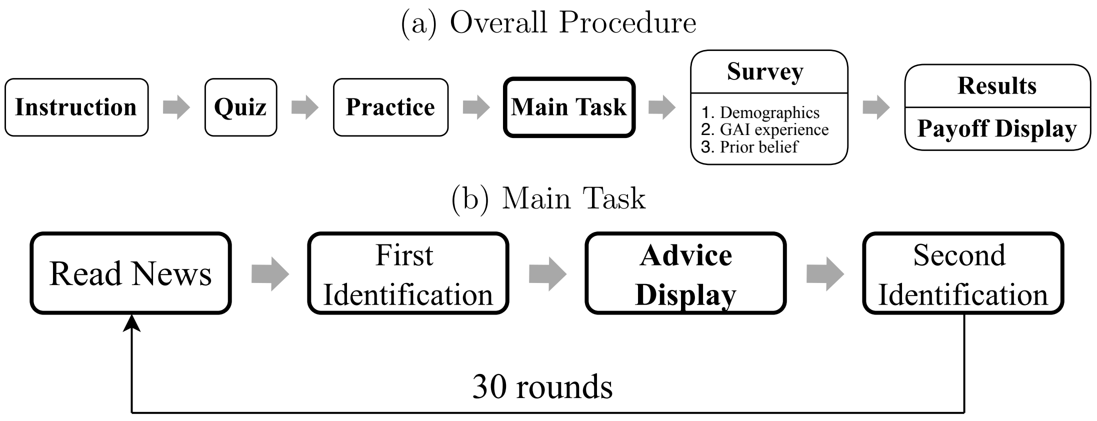


Figure 1: Experimental Procedure: (a) Overall Procedure and (b) Main Task

comparisons across AI, peers, and experts remain scarce, particularly in controlled settings that allow clean identification of source effects.

Third, we adopt the JAS paradigm using the WOA metric and further examine the advice-taking process by decomposing reliance using activation-integration model, providing additional evidence on the mechanisms through which AI advice influences decision making.

3 Experimental Design

3.1 Procedure

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

In the experiment, after reading through the instructions (See an English translation in Online Appendix F), each participant was asked to take a quiz (see Online Appendix G) to ensure they understood the rules. Then, they practiced once and entered the main tasks. After finishing all the tasks, they were asked to complete some survey questions, and the results and the final payoff were shown.

Table 1: News Materials

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

3.2 Main Task

The main task consists of 30 rounds of a deepfake detection task implemented within the JAS framework, as illustrated in Panel (b) of Figure 1. Representative experimental screens are provided in Online Appendix I.1.

There are four stages in each round. Participants first read a deepfake news article and report their initial identification of the proportion of human-written content in the article ($HMpro$). They then receive advice from ChatGPT or Human peers. After receiving the advice, participants are asked to submit a second identification. No time constraints are imposed, except for a 10-second Advice Display stage.

Details of the deepfake news materials, the deepfake detection task, treatments and two identifications are described below.

3.2.1 Deepfake News Materials and Deepfake Detection

The deepfake news articles were Japanese deepfake news collected from an open deepfake news dataset.[3] We randomly selected 30 news articles, primarily covering topics such as politics, sports, meteorology, and public safety.[4] The news items fall into three types, as summarized in Table 1, where "Length" refers to the number of characters.

The totally real news articles were written by humans and collected from Japanese Wikinews,[5] the totally fake news articles were generated by OpenAI's Japanese GPT-2 model, and the partially fake news articles consisted of both real and fake content. In

[3] https://github.com/tanreinama/japanese-fakenews-dataset?tab=readme-ov-file

[4] The original Japanese texts are available upon request. Detailed category definitions and short descriptions of the news content are provided in Online Appendix E.3.

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

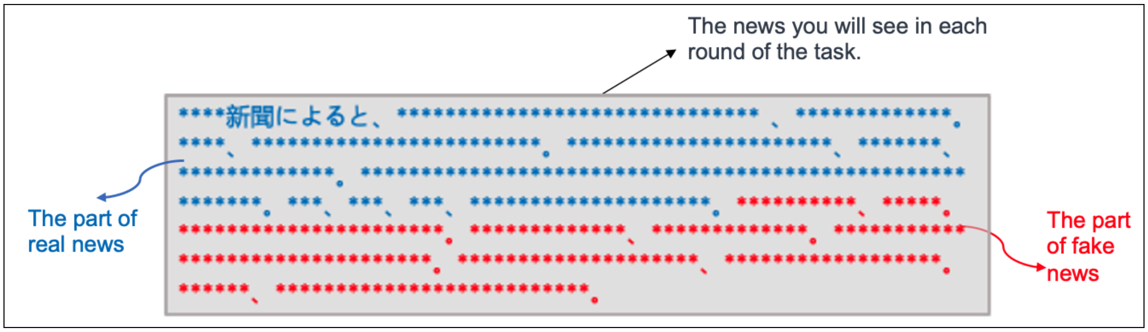


Figure 2: Instructional Illustration of News Composition Shown to Participants

the partially fake news, the first part of the article was human-written and the second part was AI-generated; the human-written part always appeared first.

Figure 2 illustrates the explanatory guidance provided to participants in the experimental instructions regarding the composition of news articles. The proportion of the "real" (human-written) part, denoted as $HMpro$, is defined as

$$HMpro_r = \frac{\text{the length of human-written part of the News in Round } r}{\text{the length of the News in Round } r} \times 100,$$

where $HMpro_r = 0$ represents totally fake news, $HMpro_r = 100$ represents totally real news, and $HMpro_r \in (0, 100)$ represents partially fake news in round r.

The **deepfake detection task** in this study is therefore defined as follows: *participants read a deepfake news article and identify its* $HMpro$ (the proportion of the "real" (human-written) part), which is operationally interpreted as the degree of "authenticity." In this context, "authenticity" serves purely as an operational label for $HMpro$ and does not refer to the factual truthfulness of the news content.

In the instructions, participants are clearly informed about the composition style of the news materials, along with the definition of $HMpro$. They are also informed that $HMpro$ represents the degree of "authenticity" they must identify and report. The true $HMpro$ of 30 pieces of news ($HMpro^*$) were assigned in a random sequence as shown in Figure 3.

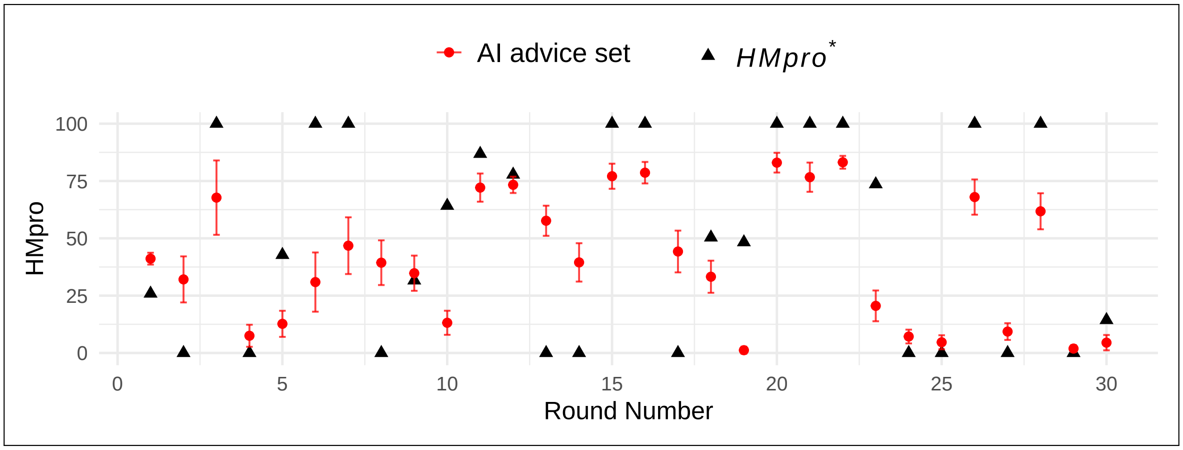


Figure 3: $HMpro^*$, the original AI advice set and Round Number

Note: The points marked with "▲" represent the true $HMpro$ values in the 30-round tasks ($HMpro^*$). 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

After each participant had read the news, they were asked to report a number between 0 and 100 to represent their first identification of $HMpro$ with a slider (see Figure I.2 in Online Appendix I.1). 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 Treatments by Advice Source: AI vs. Human

The Advice Display Stage (see Figures I.3–I.6 in Online Appendix I.1) follows a *between-subject* design with two treatments that differ only in the source of advice: an **AI** treatment and a **Human** treatment. The details are as follows,

AI treatment. The advice was one response randomly selected from 24 responses generated by ChatGPT using the GPT-4 model before the experiment, using the following prompt:

> – *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.*

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.

Human treatment. The advice was randomly selected from another participant's first identification (*Initial Response*) within the same treatment.

In each round, advice is independently randomized at the participant level, such that participants may receive different advice even when evaluating the same news article. This design helps mitigate potential biases arising from participants' beliefs about being observed by others, including the spotlight effect (Gilovich et al., 2000).

3.2.4 Second Identification

In the final stage, participants submit a second identification (*Final Response*) of *HMpro*. Both the participant's initial response and the advice for that round are displayed on the slider using distinct colors (see Figures I.7–I.10 in Online Appendix I.1).

3.3 Survey Questions About Prior Beliefs

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

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," which do you think can provide more accurate responses?*

3.4 Final Payoff

The participants' final payoff consists of a fixed participation fee of 500 JPY and an additional performance–based payoff. Specifically, the additional payoff, π, was calculated based on the accuracy of one randomly selected identification from all their identifications throughout the experiment (a total of 30 rounds × 2 identifications = 60 identifications), determined using the following quadratic equation:

$$\pi = \max\{0,\ 2300 - 0.3 \times (R_{rd} - HMpro^{*}_{rd})^2\}\ \text{JPY},$$

where R_{rd} is the randomly selected identification, and $HMpro^{*}_{rd}$ denotes the corresponding true $HMpro$ of the deepfake news in the selected round, after rounding.

3.5 Materials and Summary

The main experiment was conducted in the laboratory at the Institute of Social and Economic Research (ISER) at the University of Osaka on November 7 and 9, 2023, and October 28 and 29, 2024. We recruited 87 participants who were students at the University of Osaka registered in the ORSEE (Greiner, 2015) database of ISER. All participants were native Japanese speakers, 42 out of whom were assigned to the **Human** treatment and 45 to the **AI** treatment.[6] 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. Variable definitions are presented in Table 2, and comparisons of demographic data are illustrated in Figure E.1 in Online Appendix E.1.

During the experiment, participants were prohibited from using any of their own electronic devices, including smartphones and tablets. Although they completed the

[6] 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 treatment.

Table 2: Demographic Statistics

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

tasks on the laboratory's computers, Internet connectivity within the experiment software was also disabled.

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

3.6 Hypotheses

As discussed in the **Section** 2, deepfakes pose substantial societal risks, yet both human and machine-based detection—including AI detectors—remain imperfect and can generate secondary harms. At the same time, reliance on AI tools is increasingly observed across domains, but its implications remain context-dependent and not yet well understood. Despite these concerns, there is little systematic evidence on how individuals rely on AI-based advice when the task itself is to detect deepfake content. In such settings, excessive reliance on imperfect detection tools may fail to mitigate deepfake harms and may even introduce additional risks.

Against this backdrop, we begin by examining whether individuals rely more on

AI-based advice than on advice from human peers when detecting deepfake news. This comparison provides a natural and testable benchmark for understanding the emergence of AI reliance in a context marked by growing societal concern over deepfakes.

Accordingly, we propose the following hypothesis:

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

As noted, the news materials used in the tasks consist of three types of news—totally real, partially fake, and totally fake—with the proportion of AI-generated content ranging from 0% to 100%. This design reflects real-world deepfakes, which often combine AI-generated content with human editing rather than appearing as purely synthetic or purely human-made.

Prior work shows that ambiguous content is more difficult for individuals to evaluate than extreme cases (Friggeri et al., 2014; Lewandowsky et al., 2017). Consistent with this view, in the context of deepfake detection, estimating a specific proportion of human-written content ($HMpro \in [0, 100]$) is substantially more challenging than making a binary judgment about whether an article is totally human-written or totally AI-generated ($HMpro \in \{0, 100\}$). Consequently, when faced with news that contains a non-extreme proportion of AI-generated content, participants may experience greater uncertainty and become more inclined to rely on external advice.

This consideration leads to the following hypothesis:

H2: The reliance level is higher when the news is partially fake than when it is totally fake or totally real.

4 Results of Main Experiment

This section presents the main findings of the experiment. We first examine participants' performance in detecting deepfake news. We then analyze their reliance on ex-

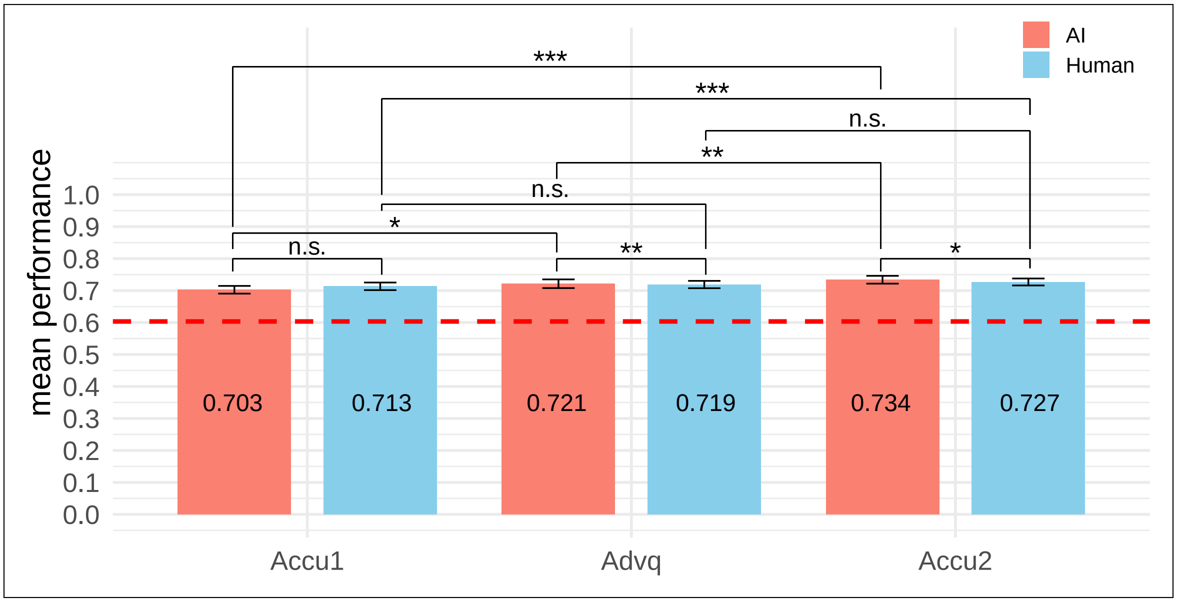


Figure 4: Overall Performance

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. $Accu_1$, $Advq$, and $Accu_2$ are compared within treatments using the Wilcoxon signed-rank test, and between treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method. The red dashed line indicates the mean accuracy ($= 0.603$) of an *uninformative baseline*, which predicts a fixed value of 50 for every round.

ternal advice and the corresponding treatment effects. Finally, we investigate whether the proportion of human-written content within each news article affects participants' performance and their reliance on advice.

4.1 Performance

4.1.1 Basic Comparisons

We first investigated participants' performance in the deepfake detection task in our experimental setting by calculating their accuracy as

$$Accu^{i}_{k,r} = 1 - \frac{|Response_{k,r} - HMpro^{*}_{r}|}{100},$$

where $k \in \{1, 2\}$, and $Accu^{i}_{1,r}$ and $Accu^{i}_{2,r}$ are participant i's accuracy for "*Initial Response*"

($Response^i_{1,r}$) and "*Final Response*" ($Response^i_{2,r}$) in round r, respectively. $HMpro^*_r$ is the true $HMpro$ of the news article in round r.

Similarly, we defined the "advice quality" ($Advq$) for participant i in round r as the accuracy of the advice:

$$Advq^i_r = 1 - \frac{|Advice^i_r - HMpro^*_r|}{100}$$

Importantly, although this measure is constructed using the true value $HMpro^*_r$, participants do not observe this benchmark. Thus, $Advq$ is interpreted as a researcher-side proxy for advice quality, which may shape participants' perceived advice quality.

Figure 4 reports the mean $Accu_1$, $Accu_2$ and mean $Advq$. Participants' initial accuracy in detecting deepfake news—before receiving any advice—was 70.8% on average, significantly above the *uninformative baseline* (Wilcoxon signed-rank test, Holm-Adjusted $p < 0.001$), indicating that the task is tractable and that participants were not responding randomly. The average quality of AI advice was 0.721, compared with 0.719 for advice from human peers. Although the difference in means is small, the Mann–Whitney U test shows a statistically significant difference (Holm-Adjusted $p = 0.0022$), suggesting that **AI can provide more accurate advice than Human peers**.

Initial accuracy did not differ across treatments. However, final accuracy was significantly higher under the AI treatment than under the Human treatment (Holm-Adjusted $p = 0.0114$). As with advice quality, this difference appears to reflect a broader distributional shift: participants' post-advice accuracy tends to be higher when the advice comes from AI, even though the difference in mean accuracy remains modest.

OLS regression results are presented in Table A.1 in Online Appendix A. The coefficient on the treatment indicator Tai, which takes value 1 for AI treatment, is not statistically significant in most specifications. In contrast, the coefficient on $Advq$ is positive and statistically significant, indicating that the higher performance observed

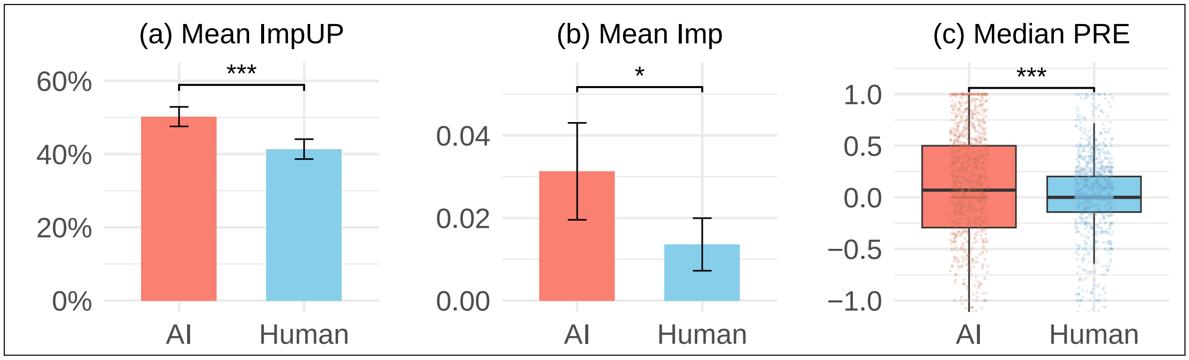


Figure 5: Performance Improvement

Note: ImpUP is compared across treatments using Fisher's exact test, and Imp and PRE are compared across treatments using the Mann–Whitney U test. For PRE, 124 observations are excluded because participants with $Accu_1 = 1$ have undefined PRE values. $^+ p < 0.1$, $* \ p < 0.05$, $** \ p < 0.01$, $*** \ p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars in (a) and (b) denote 95% confidence intervals across participants. In panel (c), each point corresponds to a participant–round PRE observation; boxes represent interquartile ranges with median lines, and whiskers extend to 1.5 times the interquartile range. The y-axis is restricted to the range $[-1, 1]$ to highlight the central distribution.

in the AI treatment is primarily driven by advice quality rather than the identity of the advice source.

4.1.2 Performance Improvement

Figure 4 shows that, participants in both conditions improved after receiving external advice and revising their initial identifications, especially in AI treatment, which suggest that AI advice provides more effective support in detecting deepfake news than advice from human peers. We then plot the *indicator for performance improvement* (ImpUP = 1 if $Accu_2 > Accu_1$, and 0 otherwise), the *raw performance improvement* (Imp = $Accu_2 - Accu_1$), and the *proportional reduction in error* (PRE)[7] in Figure 5.

Across all three measures, participants in the AI treatment perform better than those in the Human treatment: *they are more likely to improve their accuracy, achieve*

[7] For round r, $\text{PRE}_r = \dfrac{\text{Error}_{1,r} - \text{Error}_{2,r}}{\text{Error}_{1,r}} = \dfrac{(1 - Accu_{1,r}) - (1 - Accu_{2,r})}{1 - Accu_{1,r}} = \dfrac{Accu_{2,r} - Accu_{1,r}}{1 - Accu_{1,r}} \in (-\infty, 1]$, which measures the *fraction of the initial error removed* by the second identification: PRE = 1 means the initial error is fully eliminated; PRE = 0 means no change; PRE < 0 indicates deterioration. PRE is computed when $Accu_{1,r} < 1$.

larger improvements, and eliminate more initial errors after receiving AI advice.

We further analyze these patterns by estimating Probit regressions for ImpUP and OLS regressions for Imp. The results are reported in Tables A.2 and A.3 in Online Appendix A.

Across both the Probit specification (ImpUP) and the OLS specification (Imp), advice quality ($Advq$) emerges as the strongest and most consistent predictor of performance improvement. **Higher-quality advice substantially increases both the likelihood and the magnitude of improvement**. On average, **AI advice generates slightly larger improvements than human advice**. When we include the interaction term $Tai \times Advq$, the pattern becomes clearer. The negative coefficient on Tai, combined with the large positive coefficient on the interaction term, indicates that **the effectiveness of AI advice is highly dependent on advice quality**. *When advice quality is low, AI advice is less likely than human advice to produce performance gains. As advice quality increases, however, the relative effectiveness of AI advice rises rapidly.* **At sufficiently high levels of advice quality**, AI advice becomes substantially more effective than human advice in improving participants' performance.

Result 1. *AI improves participants' performance primarily because it provides higher-quality advice.*

4.2 Reliance Level

The degree of "reliance on external advice source" is measured by the "weight of advice" (Önkal et al., 2009), which is defined for participant i in the task of round r, as follows.

$$WOA_r^i = \frac{Final\ Response_r^i - Initial\ Response_r^i}{Advice_r^i - Initial\ Response_r^i}$$

This quantification approach produces a continuous outcome, where a higher WOA indicates a greater reliance on external advice. Additionally, a WOA greater than 0.5

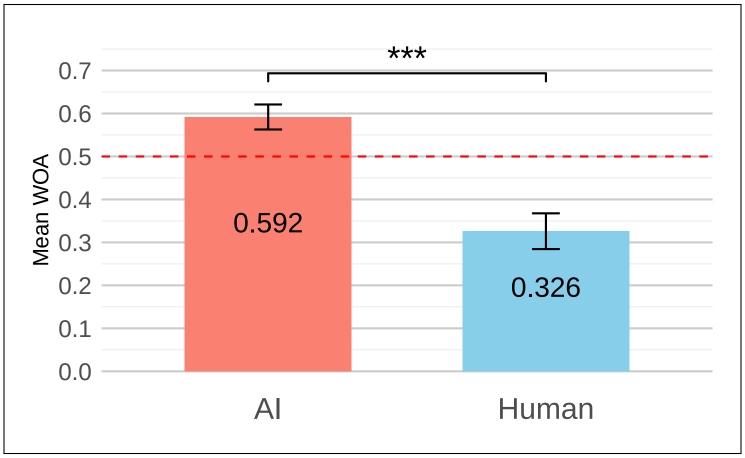


Figure 6: WOA Across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. WOA is compared across treatments using the Mann–Whitney U test. The red dashed line ($WOA = 0.5$) represents the baseline in which participants place equal weight on the advice and on their initial response.

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. In contrast, a WOA less than 0.5 signifies a lower degree of reliance on external advice.

In our analysis of the main experiment, we excluded 116 observations where WOA was undefined (i.e., when the initial response was equal to the advice given) and kept all negative values, resulting in an adjusted sample size of 2494 (1291 in the **AI** treatment and 1203 in the **Human** treatment).

Figure 6 shows the mean WOA for the two treatments in the main experiment. The average WOA was 0.592 in the **AI** treatment and 0.326 in the **Human** treatment, with the former being significantly higher. Both treatment means also differ significantly from the benchmark value of 0.5 (Mann–Whitney U test, Holm-Adjusted $p < 0.001$).

These results indicate that participants in the **Human** treatment placed relatively little weight on advice from their peers, tending instead to adhere to their initial judgments. In contrast, participants in the **AI** treatment placed substantially more weight on advice from ChatGPT and were more willing to adjust their initial responses accordingly, providing strong support for **H1**.

Result 2. *People rely on ChatGPT more than their peers when detecting AI-generated content in deepfake news articles*

4.3 The Role of Human-written Proportion

In addition to the categorical news types described in Table 2, we examine whether variation in the proportion of human-written (non-AI-generated) content[8] within each article influences participants' detection performance and their reliance on external advice, enabling a more fine-grained analysis of behavior in deepfake detection tasks.

Performance. Tables A.1–A.3 in Online Appendix A report the regression results examining the role of human-written proportion in shaping performance outcomes. In Table A.1 (OLS regressions of performance), the coefficients on $isreal$ and $isfake$—indicators for totally human-written and totally AI-generated articles—are not statistically significant, and they remain insignificant in Table A.2 (Probit regressions of performance improvement). In contrast, both coefficients become significantly positive in Table A.3 (OLS regressions of performance improvement), indicating that participants improve more when evaluating extreme cases (totally human-written or totally AI-generated articles) than when evaluating mixed-content articles. The coefficient on $HMpro$ is significantly negative in Table A.1, although its magnitude is negligible, and it is not significant in Tables A.2 or A.3. This pattern suggests that although participants' initial performance decreases slightly as articles become more human-written, the proportion of human-written content does not systematically affect whether participants improve or by how much they improve after receiving advice.

[8] Note that the proportion of human-written content is mechanically equivalent to the complement of the AI-generated proportion; higher values of one necessarily imply lower values of the other. We adopt the former throughout for consistency with the main text.

Result 3. *The proportion of human-written content within a deepfake news article does not exert a systematic effect on participants' detection performance or on their improvement after receiving advice.*

Reliance level. Table A.4 in Online Appendix A reports the OLS regression results for reliance measured by WOA. Across all specifications, the coefficients on $HMpro$, $isreal$, $isfake$, and their interactions with the treatment indicator are not statistically significant. These findings indicate that the extent of AI generation in a deepfake news does not appear to influence reliance behavior in the deepfake detection task. Therefore, our second hypothesis, **H2**, is rejected.

Result 4. *Participants' reliance on external advice does not vary with the proportion of human-written content in the deepfake news article.*

5 Additional Experiments

5.1 New treatments and Hypotheses

As a robustness check, we conducted three additional sessions on June 16, June 30, and October 8, 2025, in the laboratory of the ISER at the University of Osaka. Relative to the main experiment, these sessions introduced two new advice sources and moved the post–main-task survey to an earlier stage, positioned between the Quiz and Practice sections. The three additional treatments are summarized below.

1. **Expert**. The survey was administered immediately after the Quiz and before the Practice stage. In the Main Task, participants received advice given by linguistic experts.[9]

[9] Participants were informed that each piece of advice came from an identification made by *a linguistic expert*—professors, assistant professors, or lecturers specializing in linguistics. We collected Experts' advice via a separate questionnaire from 11 experts affiliated with the University of Osaka, Kwansei Gakuin University, and Keio University, who provided their own identification of at least five deepfake news articles used in the main experiment, without online search. Note that unlike hu-

2. **preHuman**. The survey was administered before the Practice stage, same as in the **Expert** treatment instead of after the main task. In the Main Task, the advice shown to participants was a randomly selected first identification reported by *participants in the* ***Human*** *condition of the main experiment reported above.*[10]

3. **AIadd**. The survey was administered immediately after the Quiz and before the Practice stage. All other procedures replicated the **AI** treatment in the main experiment.

The purpose of the **Expert** treatment is to extend the comparison beyond ChatGPT and student peers by introducing advice generated by linguistic experts. The **preHuman** treatment is designed to examine whether the reliance on advice from human peers in the main experiment may have been affected by social interactions or other-regarding preferences.[11] The **AIadd** treatment is intended to test whether the timing of the survey influences participants' prior beliefs about the advice source.[12]

Previous research suggests that people tend to trust responses generated by LLMs more than those provided by human experts (Shekar et al., 2025). We also assume that participants generally do not distinguish between advice coming from peers in the same session and peers who participated earlier. Therefore, for the two newly added treatments, we have the following hypotheses:

H3: Compared to linguistic experts, people tend to rely more on ChatGPT to detect deepfake news.

H4: Compared to human peers (students who participated in this experiment before), people tend to rely more on ChatGPT to detect deepfake news.

man peer's initial evaluation, these experts were not given monetary incentive for providing accurate evaluation. For each of the 30 articles, we collected at least three expert identifications.

[10] Participants were informed that each piece of advice came from an first identification made by a participant in the experiment conducted in the previous year.

[11] We thank anonymous reviewer for pointing out this possibility.

[12] The **Expert** and **preHuman** treatments were preregistered at aspredicted.org (#233371). The **AIadd** treatment was not preregistered because the main task remained unchanged.

5.2 Results of Additional Experiment

We recruited 133 participants who were also students at the University of Osaka registered in the ORSEE (Greiner, 2015) database of ISER. All participants were native Japanese speakers, 44 out of whom were assigned to the **Expert** treatment, 47 to the **preHuman** treatment and 42 to the **AIadd** treatment. Participants in the **Expert** treatment earned an average final payoff of 2420 JPY, while those in the **preHuman** treatment earned 2425 JPY, and those in the **AIadd** treatment earned 2513 JPY. Comparisons of demographic data are illustrated in Figure E.1 in Online Appendix E.1.

5.2.1 Advice Quality

Figure 7 reports the quality of advice shown to participants across **AI**, **Human**, **preHuman** and **AIadd** treatments, which shows that ChatGPT achieves higher accuracy than human participants in the deepfake detection task, and therefore can provide higher-quality advice.

Here, we exclude the **Expert** treatment from *Advq* comparisons because experts' advice was collected under a very different condition from human advice (in particular, it was not incentivized for accurate evaluation); therefore, we refrain from presenting and comparing their accuracy with other advice.[13] Nevertheless, expert advice's *Advq* is included in subsequent analyses that examine the relationship between *Advq* and participants' reliance and detection performance. This is because participants were not informed about the incentive structure behind expert advice, and thus any differences in incentive provision do not affect participants' beliefs or decision-making conditional on the realized *Advq*.

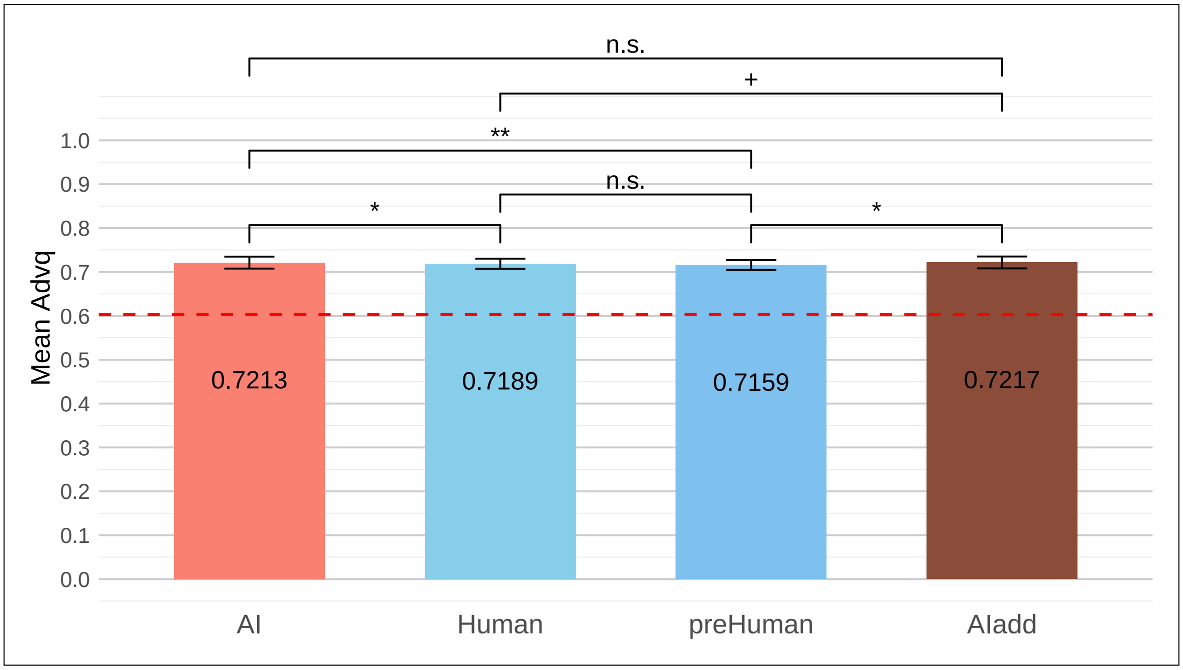


Figure 7: Mean *Advq* Across Treatments (Excluding **Expert**)

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. The Expert treatment is excluded from this figure because expert advice was collected via an unincentivized questionnaire and is therefore not comparable in advice quality to other treatments. *Advq* is compared across treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method. The red dashed line indicates the mean accuracy (= 0.603) of an *uninformative baseline*, which predicts a fixed value of 50 for every round.

5.2.2 Performance

Initial Detection Accuracy. Figure 8 shows that $Accu_1$ does not differ systematically across treatments. Columns (1)–(3) of Table B.1 in Online Appendix B further indicate that the human-written proportion itself has no meaningful effect on baseline detection accuracy. The indicators for extreme cases—articles that are totally human-written ($isreal$) or totally AI-generated ($isfake$)—are statistically significant, but their magnitudes are small: the coefficients imply only a 3%–4% point reduction in accuracy. This pattern suggests that participants behave somewhat cautiously—and may hedge against potential errors—when facing boundary cases in the absence of advice, though

[13] For reference only, the mean value of *Advq* for expert advice is 0.6919.

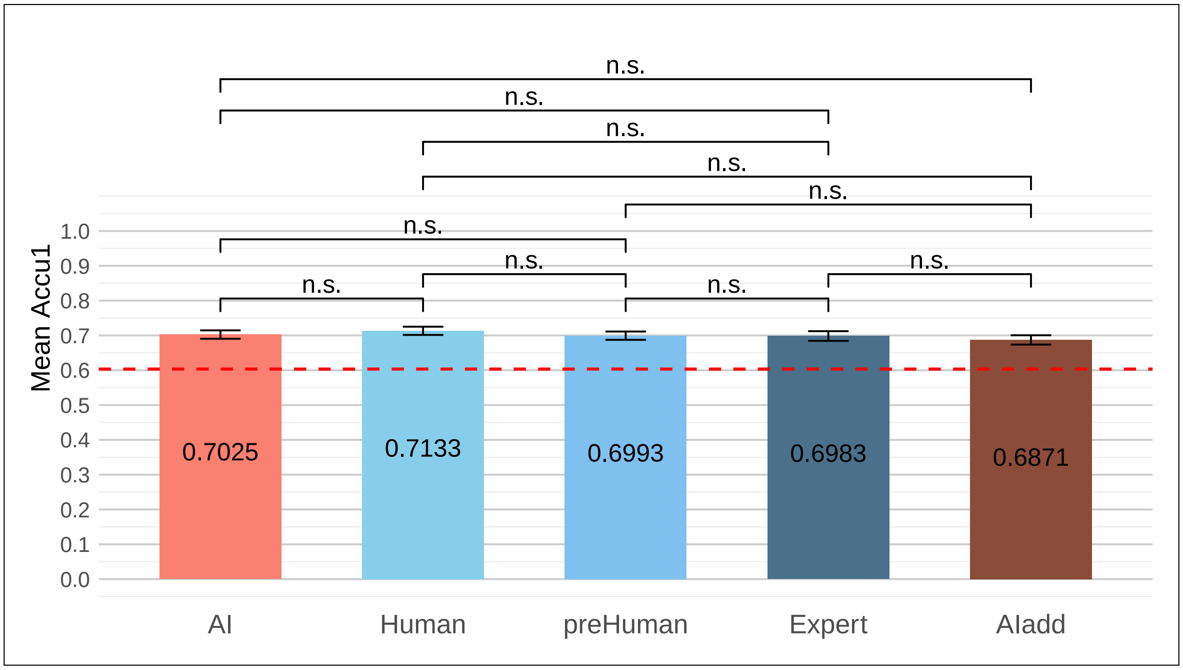


Figure 8: Mean $Accu_1$ Across Treatments

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. $Accu_1$ is compared across treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method.

the practical impact on baseline performance remains limited.

Final Detection Accuracy. Figure 9 reports the average $Accu_2$ across the five treatments. The **AI** treatment achieves the highest final accuracy. The **Expert** treatment also attains a higher $Accu_2$ than both the **Human** and **preHuman** treatments. The difference between the **AI** and **Expert** treatments is not statistically significant.

Columns (4)–(6) of Table B.1 and the corresponding interaction results in Table B.2 in the Online Appendix B show that the human-written proportion ($HMpro$) has a statistically significant but substantively negligible effect on $Accu_2$: although the coefficient is negative and significant, its magnitude (about –0.0003) implies an effect that is economically minimal.

By contrast, advice quality ($Advq$) plays a dominant role. A one percentage point increase in advice quality (i.e., a 0.01 increase on the 0–1 scale) raises final accuracy

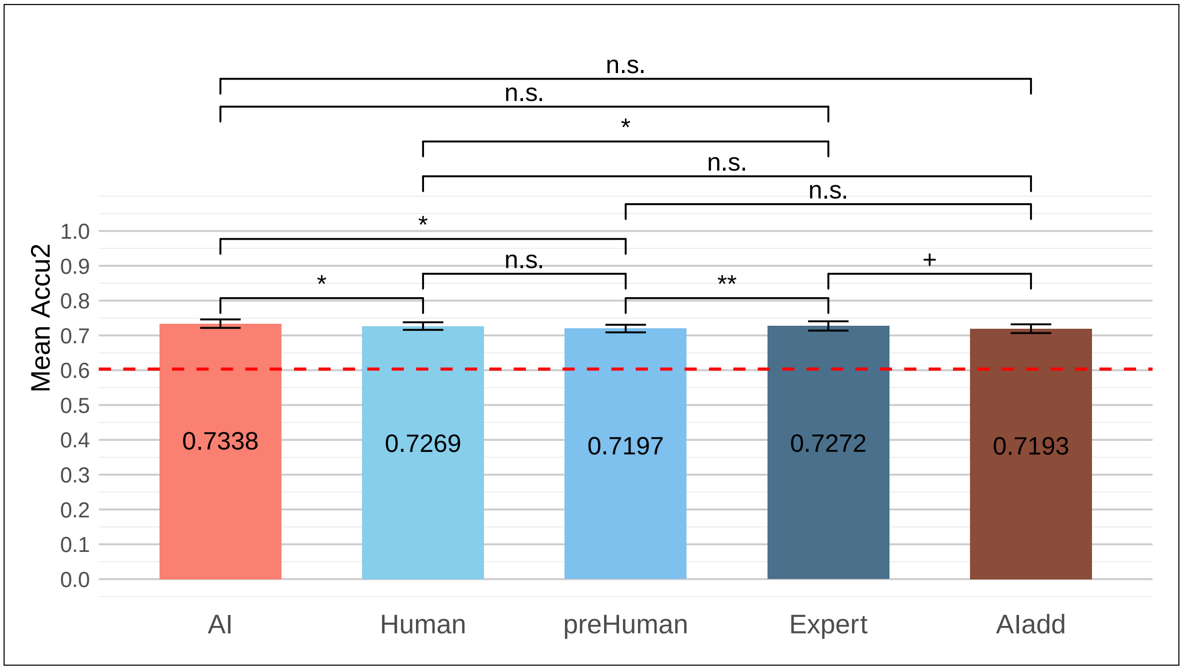


Figure 9: Mean $Accu_2$ Across Treatments

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. $Accu_2$ is compared across treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method. The red dashed line indicates the mean accuracy (= 0.603) of an *uninformative baseline*, which predicts a fixed value of 50 for every round.

by roughly 0.0055 in the baseline specification. This effect is even stronger in the AI treatment: the interaction term (0.223) implies that the total marginal effect of advice quality for AI advice is approximately $0.499 + 0.223 = 0.722$, highlighting the particularly high returns to advice quality when the advice comes from ChatGPT.

Performance Improvement. Consistent with the results from the main experiment, participants in the three additional sessions also improved their performance after receiving advice, with $Accu_2$ being significantly higher than $Accu_1$ (Wilcoxon signed-rank test, $p < 0.0001$).

Figures B.1 and B.2 in Online Appendix B further report the average Imp and ImpUP across treatments. While no statistically significant differences are observed in raw performance improvement across treatments, participants who receive advice from

AI or **Expert** exhibit a higher probability of performance improvement than those who receive advice from peer sources.

Tables B.3 and B.4 in Online Appendix B report the corresponding regression analyses for performance improvement. Across all specifications, the human-written proportion has no significant effect on the likelihood of improvement, and its effect on raw improvement—while sometimes statistically significant—is economically negligible. By contrast, advice quality ($Advq$)—particularly AI advice—exhibits a large, stable, and highly significant positive effect in every specification. Its magnitude exceeds that of treatment indicators, the human-written proportion, and the extreme-news indicators ($isfake$, $isreal$) by an order of magnitude.

This pattern indicates that performance differences across treatments are driven overwhelmingly by differences in the quality of advice participants receive, rather than by treatment identity per se. Specifically, AI advice tends to generate larger improvements because it systematically provides higher-quality guidance. Variation in $Accu_2$ and improvement therefore arises primarily because treatments expose participants to advice pools of differing quality. This supports our conclusion in Result 1: *treatment differences in the additional experiment do not operate independently of advice quality; rather, advice quality is the dominant channel through which these differences manifest.*

5.2.3 Reliance Level

Figure 10 reports the WOA values across the five treatments. The ordering of reliance is $WOA_{AI} > WOA_{Expert} > WOA_{AIadd} > WOA_{Human} \approx WOA_{preHuman}$. The comparisons between **Human** and **preHuman**, and between **AI / AIadd** and **Human / preHuman**, confirm **H4**: *reliance on peer advice in the main experiment is not driven by social interaction or other-regarding preferences. Participants rely substantially more on ChatGPT than on human peers.*

A more nuanced pattern emerges when comparing reliance across the three non-

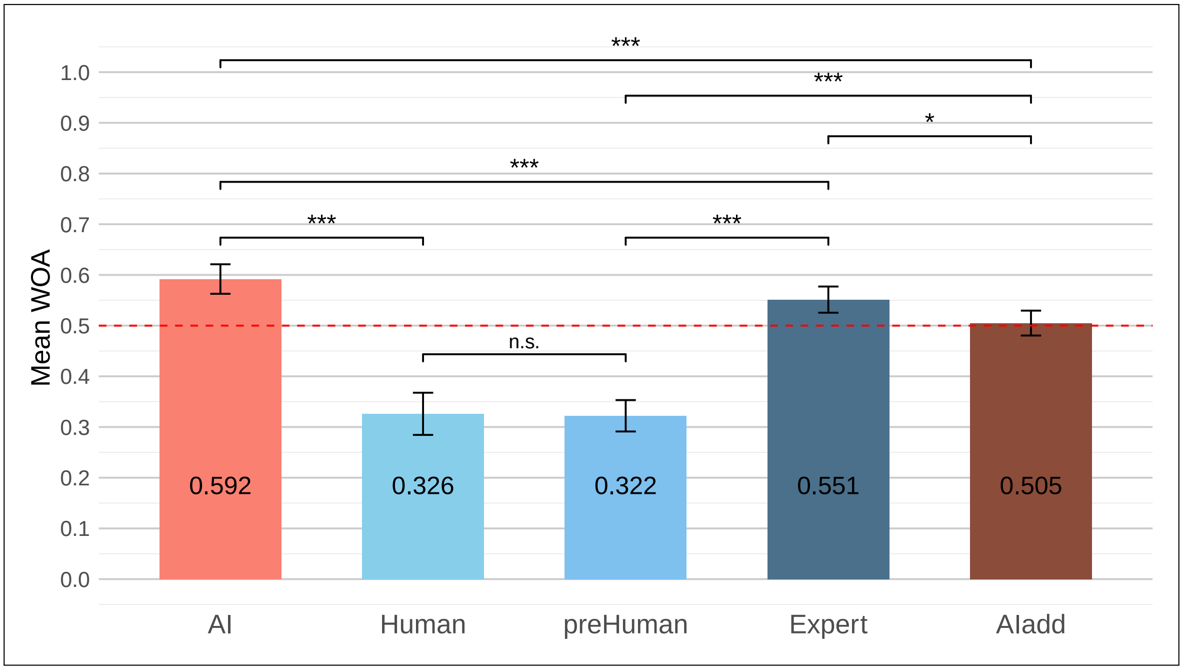


Figure 10: WOA Across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. “n.s.” means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. WOA is compared across treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method. The red dashed line ($WOA = 0.5$) represents the baseline in which participants place equal weight on the advice and on their initial response.

peer treatments. Although reliance in the **AI** treatment is significantly higher than in the **Expert** treatment (Holm-adjusted $p = 0.00002$), this comparison spans different implementation years (2023–2024 vs. 2025) and thus may reflect temporal or contextual differences rather than source credibility. By contrast, the **Expert** and **AIadd** treatments were both conducted in 2025 under an identical procedural flow, providing a cleaner within-year comparison of advice sources.[14] Reliance is significantly higher for **Expert** than for **AIadd** (Holm-adjusted $p = 0.01308$), indicating that when procedural and temporal factors are held constant, participants place more weight on advice attributed to linguistic experts than on advice attributed to ChatGPT. Accordingly, **H3** is not supported.

[14] A detailed discussion of this cross-wave divergence is provided in Online Appendix C.

Result 5. *Participants assign the greatest reliance to advice from linguistic experts, followed by advice from ChatGPT, and rely the least on advice from human peers when detecting deepfake news.*

Table B.5–B.7 in Online Appendix B report the OLS regression results for participants' reliance behavior with respect to the human-written proportion in deepfake news. Across almost all specifications, the coefficients on $HMpro$, $isreal$, $isfake$, as well as their interactions with the treatment indicators, are statistically insignificant. These results further reinforce our earlier conclusion in Result 4 that the human-written proportion of the news content does not meaningfully affect participants' reliance level.

6 Discussions

6.1 Prior Beliefs

We elicit participants' prior beliefs using three survey questions reported in Section 3.3. Since all participants answered "yes" to **SQ5**, we focus on responses to **SQ6** and **SQ7**.

SQ6 measures the self–reported frequency of using ChatGPT ($freqGPT$). Figure E.1 in Online Appendix E.1 reports the average values of $freqGPT$ across treatments.[15]

SQ7 elicits participants' beliefs about whether GAI or humans would perform better in the deepfake detection task. Importantly, this question was asked before participants observed any task outcomes or payoff information. In the two treatments of the main experiment, this question was asked after the main task but before the display of results and payoffs. In the three treatments of the additional experiment, the same question was asked before the main task began.

[15] The $freqGPT$ levels in the additional experiment are substantially higher than those in the main experiment. A likely explanation is that the main experiment was conducted in 2023–2024, whereas the additional experiment took place in 2025, a period during which public familiarity with and usage of AI tools increased markedly.

Based on responses to **SQ7**, we constructed a dummy variable, *prefAdvSrc*, capturing “participants’ relative preference for the offered advice source,” or “prior beliefs about relative advice effectiveness”, coded as:

$$prefAdvSrc_i = \begin{cases} 1, & \text{if } i \text{ in AI/AIadd treatment} \\ & \quad \text{and trusted AI performs better,} \\ 1, & \text{if } i \text{ in Human/preHuman/Expert treatment} \\ & \quad \text{and trusted Human performs better,} \\ 0, & \text{otherwise.} \end{cases}$$

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

Figure 11 presents the *prefAdvSrc* across treatments. Among all pairwise comparisons, only the difference between the **Human** and **preHuman** treatments is statistically significant; all other comparisons show no significant differences. In particular, we do not detect a statistically significant difference in *prefAdvSrc* between the **AI** (fielded in 2023/2024) and **AIadd** (fielded in 2025) samples. However, because questionnaire timing and calendar time vary simultaneously across these experimental, this comparison does not cleanly identify a pure timing effect on prior beliefs; rather, it suggests that any timing- or cohort-driven differences in stated priors are not large enough to be statistically detected in our data.

Given this pattern, it is reasonable to pool observations from the main and additional experiments when analyzing *prefAdvSrc*. OLS regression results are reported in Table B.8 in Online Appendix B. The coefficient on *prefAdvSrc* is positive and statistically significant, indicating that *participants are more likely to rely on advice received from a source they prefer.* In contrast, frequent use of ChatGPT in daily life does not translate

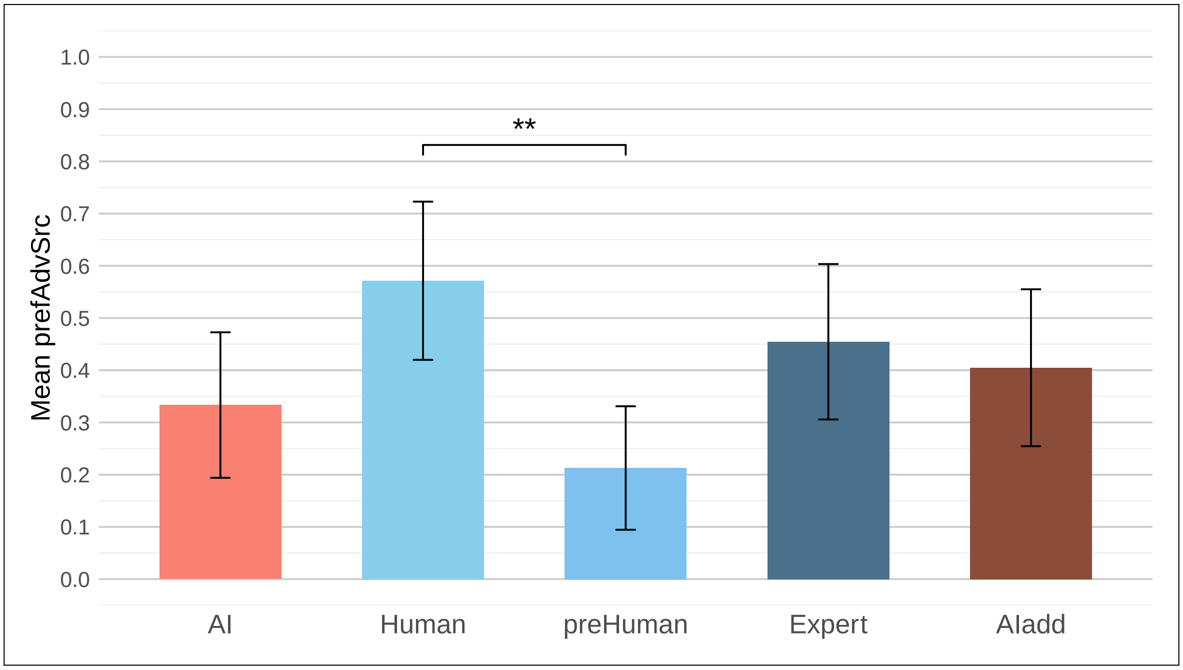


Figure 11: Mean $prefAdvSrc$ Across Treatments

Note: ${}^{+}p < 0.1$, ${}^{*}p < 0.05$, ${}^{**}p < 0.01$, ${}^{***}p < 0.001$. “n.s.” indicates that the difference is not statistically significant at the 10% level. Error bars denote 95% confidence intervals across participants. Pairwise comparisons across treatments are conducted using Fisher’s exact tests. All reported p-values are adjusted for multiple comparisons using the Holm method. None of the pairwise differences are statistically significant except for the comparison between the Human and preHuman treatments (Holm Adjusted $p = 0.0091$).

into greater reliance on AI advice in the experimental setting. In particular, participants who received ChatGPT’s advice and believed that GAI outperforms humans exhibited higher reliance on AI advice. Moreover, we find little evidence that differences in the preference-elicitation protocol materially alter stated preferences or their association with reliance. Specifically, the interaction patterns between $prefAdvSrc$ and the AI-based treatments are qualitatively similar in **AI** and **AIadd**. This stability supports pooling the main and additional experiments for the purpose of analyzing how $prefAdvSrc$ relates to reliance behavior, while acknowledging that calendar time and elicitation timing vary jointly across experimental waves.

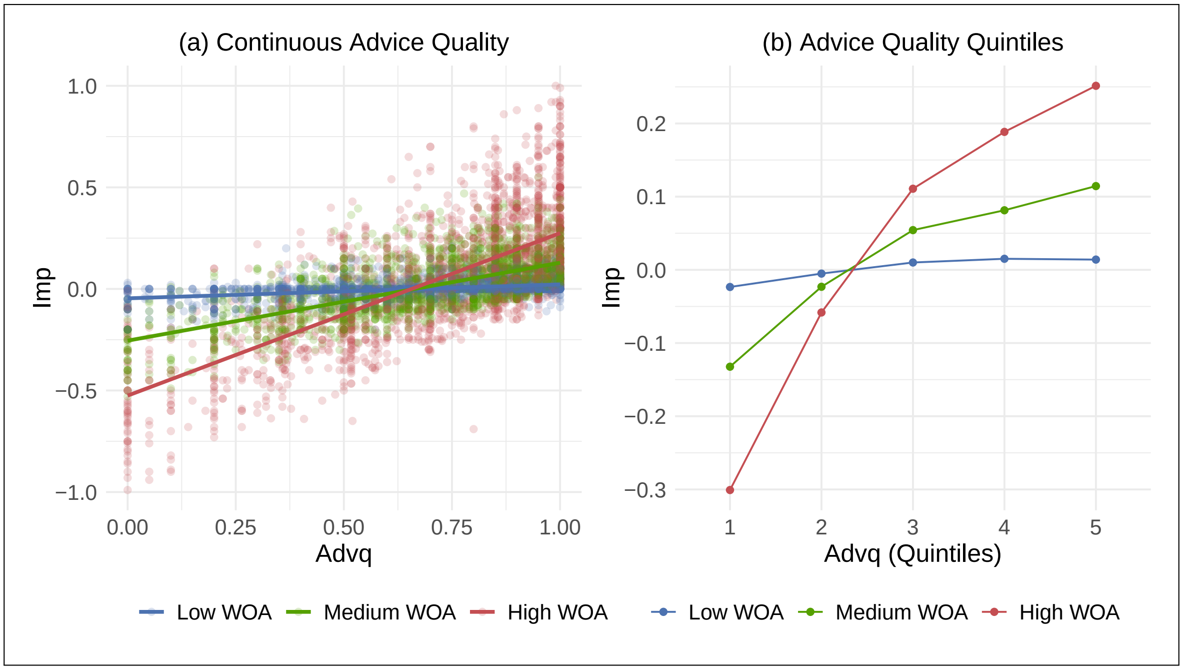


Figure 12: Advice Quality, WOA & Performance Improvement

Note: All observations are classified into low, medium, and high WOA groups based on terciles of the WOA distribution. Panel (a) plots the relationship between continuous advice quality and performance improvement, with fitted linear trends shown separately for each WOA group. Panel (b) provides a non-parametric illustration by grouping advice quality into quintiles and plotting mean performance improvement within each bin for the three WOA groups.

6.2 The Joint Role of Reliance and Advice Quality

To better interpret the performance results, we examine how reliance and advice quality jointly shape performance improvement. In Section 5.2.1, we show that advice quality ($Advq$) has a large and stable association with participants' performance, substantially stronger than the treatment indicators capturing advice source. The significant interaction between Tai and $Advq$ in Tables B.2 and B.4 in Online Appendix B further suggests that, conditional on advice quality, participants receiving AI advice are more likely to translate advice into performance improvements than those receiving peer advice.

These findings motivate a simple distinction: *advice source mainly shapes reliance behavior (WOA)—the weight participants place on advice—whereas advice quality determines how beneficial advice is conditional on being taken.* Figure 12 visualizes this mechanism. The relationship between advice quality and performance improvement is weak among participants with low WOA, but becomes progressively stronger for those with medium and high WOA, with the steepest gradient among the high-WOA group.[16]

Importantly, participants do not observe objective advice quality at the time of decision making. Any association between ex post advice quality and their behavior must therefore operate through participants' perceived quality when evaluating the advice. Thus, the pattern in Figure 12 implies that ex post advice quality translates into performance improvements primarily when participants choose to rely on the advice, consistent with a gating role of reliance.

At the same time, high reliance should not be interpreted as normative superiority of a given advice source. In this sense, **reliance governs whether advice can affect performance, while advice quality determines whether such reliance is ultimately beneficial**.

Supporting this interpretation, Tables B.9 and B.10 in the Online Appendix B show that higher ex post advice quality[17] is positively correlated with reliance behavior. **Although participants do not observe objective advice quality at the time of decision making, this correlation suggests that their subjective evaluation of advice quality is, on average, aligned with realized advice quality.** We interpret this correlation as suggesting that objective advice quality is positively associated with participants' perceived advice quality when evaluating the advice.

6.3 Decomposition of Reliance

Our main analysis relies on the WOA to quantify reliance. Although widely used, WOA has a well-known limitation: when a participant's initial response coincides with the advice, the denominator becomes zero, producing undefined or extreme values that

[16] An OLS regression of Imp on $Advq$, WOA, and $Advq \times$ WOA yields a positive and significant interaction ($\beta = 0.513$, $p < 0.001$; participant-clustered standard errors).

[17] Measured both contemporaneously ($Advq$) and with a one-period lag ($AdvqLag$).

obscure the underlying behavioral process. Rather than clipping these outliers, we exclude only the undefined cases (4.4% of observations). However, this approach still prevents us from using the full sample and ignores some extreme but behaviorally meaningful instances of reliance.

To address this issue—and to provide both a robustness check for the WOA-based analysis and a deeper examination of the advice-taking mechanism—we adopt the activation–integration framework of Vodrahalli et al. (2022), which conceptualizes reliance as a two-stage process: whether participants become activated to use the advice, and, conditional on activation, how strongly the advice is integrated into the final judgment. We implement this framework using a Heckman selection model (Heckman, 1974, 1979).

The detailed specification and estimation results are reported in Online Appendix D; here we summarize the main findings.

Activation—*whether participants choose to take the advice*—is strongly predicted by the advice source, prior beliefs, the advice–initial gap (gap between the advice and initial response), and experienced lagged advice quality: participants are significantly more likely to take AI or Expert advice, more likely to be activated when they believe the source is effective, and more likely to be activated after both a larger advice–initial gap and recent positive experience with advice quality.

Conditional on activation, **Integration**—*the extent to which participants move toward the advice*—reflects a different set of forces. Integration is shaped by the advice source and prior beliefs, but, unlike activation, it is negatively affected by the advice–initial gap. This pattern suggests that large gaps may draw participants into considering the advice but subsequently make them more cautious about how far to adjust. Consistent with this, experienced lagged advice quality does not significantly predict integration without Heckman's correction.

6.4 Other Robustness Checks

Additional robustness checks are reported in Online Appendix E. Here we briefly summarize the main findings.

Demographics. We examine heterogeneity with respect to demographic characteristics reported in Table 2. The results show no systematic differences in either reliance or performance across demographic groups.

Decision Time. We further investigate whether decision time influences advice reliance or performance, using measures of the time spent on the first identification, the second identification, and the average reading time per character. We find no evidence of meaningful heterogeneity along these dimensions.

News Categories. We classify news articles based on topic and content into four initial categories and then aggregate them into two broader groups for balance: **HARD** news (fact-oriented and policy-relevant domains) and **SOFT** news (lifestyle, cultural, sports, and entertainment topics). Whether a deepfake belongs to the **HARD** category has no significant effect on reliance. However, participants find **HARD** deepfake news more difficult to detect. Relative to human advice, AI advice helps mitigate this difficulty, improving detection performance in **HARD** domains.

Learning Effects. Using both regression-based time controls and block-level comparisons, we find no evidence of learning in advice reliance. In contrast, performance improves over rounds as participants gain experience in detecting deepfake news. This learning effect is particularly pronounced among participants receiving AI advice, whose performance improvements are larger in the final ten rounds.

7 Conclusions

This paper studies how individuals rely on different advice sources when detecting deepfake news. We implement a laboratory deepfake-detection task in which participants identify the proportion of human-written content in synthetic news articles and receive advice from ChatGPT (GPT-4), human peers, or linguistic experts. Reliance is measured behaviorally using the WOA, and performance is evaluated by accuracy and performance improvement.

Four main findings emerge. First, participants rely more on GPT-4 than on human peers when detecting GPT-2–generated deepfake news. This difference is robust across experimental waves and persists even when peer advice is drawn from earlier sessions. Second, the human-written proportion within an article—and equivalently, the extent of AI generation—does not systematically affect performance or reliance behavior. Third, advice from ChatGPT improves performance relative to peer advice, and this advantage is explained primarily by advice quality rather than by the AI label per se: higher-quality advice generates larger performance improvements regardless of source, and the AI treatment performs better largely because it exposes participants to a higher-quality advice pool. Fourth, the relative reliance on experts versus ChatGPT is mixed across waves: Reliance on ChatGPT exceeds reliance on experts in the main experiment, whereas reliance on experts exceeds reliance on ChatGPT when both are implemented under the same additional procedural flow in 2025.

In addition to documenting treatment differences in performance and reliance, our analysis shows that participants' prior beliefs systematically shape their reliance decisions. Consistent with the activation–integration framework of Vodrahalli et al. (2022), our results further indicate that reliance unfolds sequentially, with different factors governing whether advice is taken and how strongly it is integrated. We also document heterogeneity with respect to news categories and learning effects in performance, while finding no corresponding learning effects in advice reliance.

Overall, our findings extend the study of “AI reliance” to the domain of deepfake detection and highlight the dual role of GAI—as both a potential source of misinformation and a tool for mitigating it. In settings where the object of detection is AI-generated content, reliance on AI-based advice is not necessarily detrimental and can improve performance when advice quality is sufficiently high. At the same time, the expert condition cautions that reliance may respond strongly to perceived credibility, which does not always align with accuracy. From a policy perspective, the effectiveness of AI-based detection tools depends jointly on their objective quality and on public beliefs about who—or what—is trustworthy.

Several limitations point to directions for future research. First, our deepfake materials were generated using GPT-2. As frontier models continue to produce increasingly human-like content, both detection difficulty and reliance patterns may evolve, raising the need to test with more advanced generators. Second, we employ prompted GPT-4 as the AI-based detection tool, whereas real-world users often interact with dedicated detection systems that differ in interface design, transparency, and feedback structure. Future studies could compare reliance on conversational AI with reliance on specialized detectors and examine how reliance varies across models. Third, our study is conducted in a laboratory environment. While this setting affords strict control—such as preventing participants from consulting external AI tools—it also limits sample size due to cost constraints. Field or online implementations with stronger scalability, but still with appropriate control over participants' information sources, would be a valuable complement. Fourth, we do not directly observe the strategies participants use to detect deepfake news. Because deepfake news may involve both factual inconsistencies and stylistic cues associated with AI-generated text, future work could examine the relative roles of these dimensions and how individuals trade off between them in detection tasks.

References

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

V. Agrawal, S. Kandul, M. Kneer, and M. Christen. From oecd to india: Exploring cross-cultural differences in perceived trust, responsibility and reliance of ai and human experts. *arXiv preprint arXiv:2307.15452*, 2023.

S. Ahmed and H. W. Chua. Perception and deception: Exploring individual responses to deepfakes across different modalities. *Heliyon*, 9(10), 2023.

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

A. Amin, Y. Hong, and B. Mazhar. The influence of social media deepfake images on political ideology and polarization: the mediating roles of cognitive load and confirmation bias. *Journal of Visual Literacy*, 44(3):321–339, 2025.

M. Amoozadeh, D. Daniels, D. Nam, A. Kumar, S. Chen, M. Hilton, S. Srinivasa Ragavan, and M. A. Alipour. Trust in generative ai among students: An exploratory study. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 67–73, 2024.

K. P. Arin, D. Mazrekaj, and M. Thum. Ability of detecting and willingness to share fake news. *Scientific Reports*, 13(1):7298, 2023.

P. E. Bailey, T. Leon, N. C. Ebner, A. A. Moustafa, and G. Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 42(28): 24516–24541, 2023.

J. I. Baines, R. S. Dalal, L. P. Ponce, and H.-C. Tsai. Advice from artificial intelligence: a review and practical implications. *Frontiers in Psychology*, 15:1390182, 2024.

S. Barrington, E. A. Cooper, and H. Farid. People are poorly equipped to detect ai-powered voice clones. *Scientific Reports*, 15(1):11004, 2025.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

J. Bergin. Over a dozen unis are using ai to catch ai — and getting it wrong, 10 2025. URL https://www.abc.net.au/news/2025-10-20/universities-using-ai-to-detect-students-cheating/105905804. Accessed: 2025-11-01.

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

M. Biswas and J. Murray. The influence of education and self-perceived tech savviness on ai reliance: The role of trust. In *World Congress in Computer Science, Computer Engineering & Applied Computing*, pages 3–17. Springer, 2024.

J. Y. Bo, S. Wan, and A. Anderson. To rely or not to rely? evaluating interventions for appropriate reliance on large language models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–23, 2025.

K. Boob-Engel. Is ai as trustworthy as human experts? the role of advice source, technology knowledge, and perceptions of ai's impact in organizational decision-making. *Journal of Decision Systems*, 34(1):2594620, 2025.

E. Brynjolfsson, D. Li, and L. Raymond. Generative ai at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.

A. Bussone, S. Stumpf, and D. O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*, pages 160–169. IEEE, 2015.

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

A. Chadha, V. Kumar, S. Kashyap, and M. Gupta. Deepfake: an overview. In *Proceedings of second international conference on computing, communications, and cybersecurity: IC4S 2020*, pages 557–566. Springer, 2021.

C. Chaka. Reviewing the performance of ai detection tools in differentiating between ai-generated and human-written texts: A literature and integrative hybrid review. *Journal of Applied Learning and Teaching*, 7(1):115–126, 2024.

A. Chakravarti. Oh god! open ai tool that identifies text written by chatgpt believes bible was written by ai, 2023. URL https://www.indiatoday.in/technology/news/story/oh-god-open-ai-tool-that-identifies-text-written-chatgpt-believes-bible-was-written-by-ai-2329163-2023-02-01. Updated: 2023-02-02.

C. Chauhan and G. Currie. The impact of generative artificial intelligence on research integrity in scholarly publishing. *The American journal of pathology*, 194(12):2234–2238, 2024.

C. Chen and Z. Cui. Impact of ai-assisted diagnosis on american patients' trust in and intention to seek help from health care professionals: Randomized, web-based survey experiment. *Journal of Medical Internet Research*, 27:e66083, 2025.

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

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

S.-Y. Chen, H. Kuo, and S.-H. Chang. Perceptions of chatgpt in healthcare: usefulness, trust, and risk. *Frontiers in Public Health*, 12:1457131, 2024a.

Z. Chen, C. Chen, G. Yang, X. He, X. Chi, Z. Zeng, and X. Chen. Research integrity in the era of artificial intelligence: Challenges and responses. *Medicine*, 103(27):e38811, 2024b.

M. Chevrier. *Forth essays on human perception of algorithms.* PhD thesis, Université Côte d'Azur, 2025.

A. T. Y. Chong, H. N. Chua, M. B. Jasser, and R. T. Wong. 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)*, pages 205–210. IEEE, 2023.

F. D. Davis. Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS quarterly*, 1989.

A. Diel, T. Lalgi, I. C. Schröter, K. F. MacDorman, M. Teufel, and A. Bäuerle. Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16:100538, 2024.

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

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

S. Eckhardt, N. Kühl, M. Dolata, and G. Schwabe. A survey of ai reliance. *ACM Computing Surveys*, 58(6):1–37, 2025.

B. Edwards. Why ai writing detectors don't work, 2023. URL https://arstechnica.com/information-technology/2023/07/why-ai-detectors-think-the-us-constitution-was-written-by-ai/. Explains false positives such as the U.S. Constitution being flagged as AI-generated.

A. Friggeri, L. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In *proceedings of the international AAAI conference on web and social media*, volume 8, No. 1, pages 101–110, 2014.

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

T. Gilovich, V. H. Medvec, and K. Savitsky. 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, 2000.

E. Glikson and A. W. Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, 14(2):627–660, 2020.

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

M. Groh, Z. Epstein, C. Firestone, and R. Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

M. Groh, A. Sankaranarayanan, N. Singh, D. Y. Kim, A. Lippman, and R. Picard. Human detection of political speech deepfakes across transcripts, audio, and video. *Nature communications*, 15(1):7629, 2024.

S. Gupta, U. Sharma, and L. Vardari. The influence of deepfake technology on political affairs. In *Mastering Deepfake Technology: Strategies for Ethical Management and Security*, pages 147–162. River Publishers, 2025.

V. Hallikaar. Western n.y. student's ai use accusation questions validity, raises concerns, 5 2025. URL https://spectrumlocalnews.com/nys/central-ny/news/2025/05/14/ub-student-says-false-ai-use-accusation-caused-stress--inspired-petition. Accessed: 2025-11-01.

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

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

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

K. A. Hoff and M. Bashir. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015.

M. B. E. Islam, M. Haseeb, H. Batool, N. Ahtasham, and Z. Muhammad. Ai threats to politics, elections, and democracy: a blockchain-based deepfake authenticity verification framework. *Blockchains*, 2(4):458–481, 2024.

R. Katarya and A. Lal. 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)*, pages 485–490. IEEE, 2020.

S. Kelly, S.-A. Kaye, K. M. White, and O. Oviedo-Trespalacios. What factors predict user acceptance of chatgpt for mental and physical healthcare: an extended technology acceptance model framework. *AI & SOCIETY*, pages 1–19, 2025.

H. Y. Kim and Y. S. Park. Trust dynamics in financial decision making: Behavioral responses to ai and human expert advice following structural breaks. *Behavioral Sciences*, 14(10):964, 2024.

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

G. Knilans. The dark side of ai detectors: Why accuracy is not guaranteed. https://www.tradepressservices.com/ai-detectors/, Oct. 2024. Trade Press Services blog post.

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

A. Küper, G. C. Lodde, E. Livingstone, D. Schadendorf, and N. Krämer. Psychological factors influencing appropriate reliance on ai-enabled clinical decision support systems: experimental web-based study among dermatologists. *Journal of Medical Internet Research*, 27:e58660, 2025.

C. Larkin, C. Drummond Otten, and J. Árvai. Paging dr. jarvis! will people accept advice from artificial intelligence for consequential risk management decisions? *Journal of Risk Research*, 25(4):407–422, 2022.

L. Laurier, A. Giulietta, A. Octavia, and M. Cleti. The cat and mouse game: The ongoing arms race between diffusion models and detection methods. *arXiv preprint arXiv:2410.18866*, 2024.

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

J. Q. Liu, K. T. Hui, F. Al Zoubi, Z. Z. Zhou, D. Samartzis, C. C. Yu, J. R. Chang, and A. Y. Wong. The great detectives: humans versus ai detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, 20(1):8, 2024.

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

E. Lundberg and P. Mozelius. The potential effects of deepfakes on news media and entertainment. *AI & SOCIETY*, pages 1–12, 2024.

M. A. Maggioni and D. Rossignoli. 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, 2023.

J. T. Mainz. Medical ai: is trust really the issue? *Journal of medical ethics*, 50(5): 349–350, 2024.

N. Mesbah, C. Tauchert, and P. Buxmann. Whose advice counts more–man or machine? an experimental investigation of AI-based advice utilization. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 496, 2021. doi: 10.24251/hicss.2021.496.

Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

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

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

S. Passi, S. Dhanorkar, and M. Vorvoreanu. Addressing overreliance on ai. *Handbook of Human-Centered Artificial Intelligence*, pages 1–34, 2025.

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

T. Ramluckan. Deepfakes: The legal implications. In *International Conference on Cyber Warfare and Security*, volume 19, pages 282–288. Academic Conferences International Limited, 2024.

T. R. Rebholz, A. Koop, and M. Hütter. Enhancing advice taking from generative ai through interactive conversational interfaces. 2024.

C. Reverberi, T. Rigon, A. Solari, C. Hassan, P. Cherubini, and A. Cherubini. Experimental evidence of effective human–ai collaboration in medical decision-making. *Scientific reports*, 12(1):14952, 2022.

R. Rosenbacke, Å. Melhus, M. McKee, and D. Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *Jmir Ai*, 3:e53207, 2024.

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

M. Sareen. Threats and challenges by deepfake technology. In *DeepFakes*, pages 99–113. CRC Press, 2022.

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

M. Serra-Garcia and U. Gneezy. Mistakes, overconfidence, and the effect of sharing on detecting lies. *American Economic Review*, 111(10):3160–3183, 2021.

E. Setyaningsih, H. Zainnuri, D. S. Wahyuni, and Y. Hariyanti. Efl students' use, perceptions, and reliance on chat-gpt for editing and proofreading: A technology acceptance model perspective. *JOLLT Journal of Languages and Language Teaching*, 13(3):1367–1379, 2025.

S. Shekar, P. Pataranutaporn, C. Sarabu, G. A. Cecchi, and P. Maes. People overtrust ai-generated medical advice despite low accuracy. *NEJM AI*, 2(6):AIoa2300015, 2025.

J. Sniezek and T. Buckley. Social influence in the advisor-judge relationship. In *Annual meeting of the judgment and decision making society, Atlanta, Georgia*, 1989.

K. Somoray, D. J. Miller, and M. Holmes. Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, 2025(1):1833228, 2025.

L. Sophia. The social harms of ai-generated fake news: Addressing deepfake and ai political manipulation. *Digital Society & Virtual Governance*, 1(1):72–88, 2025.

E. R. Spearing, C. I. Gile, A. L. Fogwill, T. Prike, B. Swire-Thompson, S. Lewandowsky, and U. K. Ecker. Countering ai-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*, 12(6):242148, 2025.

P. Stiefenhofer, U. Gergerlioğlu, and C. Deniz. Delegating authority to algorithms: Legitimacy and public values in european tax administrations. *Philosophy and Realistic Reflection*, 3(1):1–21, 2026.

X. Sun, R. Ma, X. Zhao, Z. Li, J. Lindqvist, A. E. Ali, and J. A. Bosch. Trusting the search: unraveling human trust in health information from google and chatgpt. *arXiv preprint arXiv:2403.09987*, 2024.

A. Tamò-Larrieux, C. Guitton, S. Mayer, and C. Lutz. Regulating for trust: Can law establish trust in artificial intelligence? *Regulation & Governance*, 18(3):780–801, 2024.

M. Thaler. The fake news effect: Experimentally identifying motivated reasoning using trust in news. *American Economic Journal: Microeconomics*, 16(2):1–38, 2024.

The Advertiser. 'robocheating' fiasco saw acu students falsely accused of using ai by an unreliable tool, 10 2025. URL https://www.adelaidenow.com.au/education/higher-education/robocheating-fiasco-saw-australian-catholic-university-students-falsely-accused-of-using-ai-by-an-unreliable-ai-tool/news-story/4a08732c84499263a709ec3bb1980802. Accessed: 2025-11-01.

The Courier-Mail. 'impossible': New ai detection tool slammed by experts, 8 2024. URL https://www.couriermail.com.au/queensland-education/impossible-new-ai-detection-tool-slammed-by-experts/news-story/c2443d79a81ff2fea705b6e8baf8377a. Accessed: 2025-11-01.

T. T. K. Tse, N. Hanaki, and B. Mao. Beware the performance of an algorithm before relying on it: Evidence from a stock price forecasting experiment. *Journal of Economic Psychology*, 102:102727, 2024.

H. M. Tun, H. A. Rahman, L. Naing, and O. A. Malik. Trust in artificial intelligence–based clinical decision support systems among health care workers: systematic review. *Journal of Medical Internet Research*, 27:e69678, 2025.

A. Uchendu, J. Lee, H. Shen, T. Le, D. Lee, et al. Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts? In *Proceedings of the AAAI*

Conference on Human Computation and Crowdsourcing, volume 11, No. 1, pages 163–174, 2023.

A. Uchendu, S. Venkatraman, T. Le, and D. Lee. 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)*, pages 1–7, 2024.

R. Umbach, N. Henry, G. F. Beard, and C. M. Berryessa. Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2024.

L. Van Bulck and P. Moons. What if your patient switches from dr. google to dr. chatgpt? a vignette-based survey of the trustworthiness, value, and danger of chatgpt-generated responses to health questions. *European Journal of Cardiovascular Nursing*, 23(1):95–98, 2024.

V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003.

O. Viberg, M. Cukurova, Y. Feldman-Maggor, G. Alexandron, S. Shirai, S. Kanemune, B. Wasson, C. Tømte, D. Spikol, M. Milrad, et al. What explains teachers' trust in ai in education across six countries? *International Journal of Artificial Intelligence in Education*, 35(3):1288–1316, 2025.

K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou. 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*, pages 763–777, 2022.

T. Walsh, N. Levy, G. Bell, A. Elliott, J. Maclaurin, I. Mareels, and F. M. Wood. *The* *effective and ethical development of artificial intelligence: an opportunity to improve our wellbeing.* Australian Council of Learned Academies, 2019.

D. Weber-Wulff, A. Anohina-Naumeca, S. Bjelobaba, T. Foltỳnek, J. Guerrero-Dib, O. Popoola, P. Šigut, and L. Waddington. Testing of detection tools for ai-generated text. *International Journal for Educational Integrity*, 19(1):1–39, 2023.

B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

A. Yamaoka-Enkerlin. Disrupting disinformation: Deepfakes and the law. *NYUJ Legis. & Pub. Pol'y*, 22:725, 2019.

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

J. Yin, K. Y. Ngiam, S. S.-L. Tan, and H. H. Teo. Designing ai-based work processes: How the timing of ai advice affects diagnostic decision making. *Management Science*, 2025.

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

Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, D. Gašević, and G. Chen. Detecting ai-generated sentences in human-ai collaborative hybrid texts: Challenges, strategies, and insights. *arXiv preprint arXiv:2403.03506*, 2024.

P. Zhang. Taking advice from chatgpt. *arXiv preprint arXiv:2305.11888*, 2023.

Y. Zhang, Y. Ma, J. Liu, X. Liu, X. Wang, and W. Lu. Detection vs. anti-detection: Is text generated by ai detectable? In *International Conference on Information*, pages 209–222. Springer, 2024a.

Z. Zhang, W. Qin, and B. A. Plummer. Machine-generated text localization. *arXiv preprint arXiv:2402.11744*, 2024b.

J. Zheng, L. Hao, K. Lu, A. Garg, M. Reese, M.-J. Yap, I.-J. Wang, X. Wu, W. Huang, J. Hoffman, et al. Do students rely on ai? analysis of student-chatgpt conversations from a field study. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 2796–2807, 2025.

Online Appendix to "Do people rely on ChatGPT more than their peers to detect deepfake news?"

Yuhao Fu[*] Nobuyuki Hanaki[†]

February 5, 2026

Contents

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

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

A Main Experiment: Regression Results

Table A.1: Baseline Treatment Effects on Performance

Dep. Var.	$Accu_1$	$Accu_1$	$Accu_1$	$Accu_2$	$Accu_2$	$Accu_2$	$Accu_2$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Tai	-0.011	-0.011	-0.021^{+}	0.005	-0.190^{***}	0.005	-0.001
	(0.014)	(0.014)	(0.012)	(0.009)	(0.037)	(0.009)	(0.008)
HMpro	-0.0003^{+}		-0.0003^{+}	-0.0003^{*}	-0.0002^{+}		-0.0003^{*}
	(0.0002)		(0.0002)	(0.0001)	(0.0001)		(0.0001)
isreal		-0.036^{*}				-0.005	
		(0.015)				(0.010)	
isfake		-0.014				0.013	
		(0.015)				(0.011)	
engr			-0.051^{***}				-0.038^{***}
			(0.015)				(0.009)
progexp			0.001				0.007
			(0.015)				(0.009)
edulevel			0.013				0.006
			(0.014)				(0.009)
Advq				0.617^{***}	0.451^{***}	0.620^{***}	0.617^{***}
				(0.028)	(0.034)	(0.028)	(0.028)
Tai×Advq					0.271^{***}		
					(0.048)		
Constant	0.731^{***}	0.730^{***}	0.757^{***}	0.296^{***}	0.414^{***}	0.278^{***}	0.313^{***}
	(0.012)	(0.009)	(0.015)	(0.022)	(0.025)	(0.022)	(0.024)
Adj. R^2	0.004	0.004	0.015	0.466	0.487	0.465	0.472
No. cluster	87	87	87	87	87	87	87
No. Obs.	2610	2610	2610	2610	2610	2610	2610

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Tai is a treatment indicator equal to 1 for participants assigned to the AI condition. $isreal$ is an indicator equal to 1 when the news is totally human-written ($HMpro = 100$). $isfake$ is an indicator equal to 1 when the news is totally AI generated ($HMpro = 0$). Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.2: Baseline Treatment Effects on Performance Improvement (Probit)

Dep. Var.	ImpUP				
	(1)	(2)	(3)	(4)	(5)
Tai	0.219^{*}	-0.477^{*}	0.219^{*}	0.255^{**}	0.243^{*}
	(0.091)	(0.215)	(0.091)	(0.089)	(0.114)
Advq	3.076^{***}	2.571^{***}	3.073^{***}	3.095^{***}	3.096^{***}
	(0.161)	(0.207)	(0.158)	(0.165)	(0.164)
HMpro	0.0002	0.0003		0.0002	0.0002
	(0.001)	(0.001)		(0.001)	(0.001)
Tai × Advq		0.919^{**}			
		(0.310)			
isreal			0.025		
			(0.079)		
isfake			0.010		
			(0.078)		
engr				0.245^{*}	0.233
				(0.102)	(0.144)
progexp				−0.051	−0.051
				(0.096)	(0.096)
edulevel				−0.001	−0.003
				(0.104)	(0.109)
Tai × engr					0.023
					(0.180)
Constant	-2.499^{***}	-2.125^{***}	-2.500^{***}	-2.633^{***}	-2.626^{***}
	(0.144)	(0.150)	(0.141)	(0.162)	(0.176)
AIC	2985.8	2977.4	2987.8	2975.1	2977
No. cluster	87	87	87	87	87
No. Obs.	2610	2610	2610	2610	2610

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.3: Baseline Treatment Effects on Performance Improvement (OLS)

Dep. Var.	Imp					
	(1)	(2)	(3)	(4)	(5)	(6)
Tai	0.017*	−0.188***	0.017*	0.011	0.020*	0.020**
	(0.008)	(0.034)	(0.008)	(0.007)	(0.009)	(0.007)
Advq	0.421***	0.246***	0.423***	0.424***	0.421***	0.422***
	(0.032)	(0.026)	(0.032)	(0.032)	(0.031)	(0.031)
HMpro	−0.00003	0.00001				−0.00003
	(0.0001)	(0.0001)				(0.0001)
Tai × Advq		0.285***				
		(0.050)				
isreal			0.024**	0.014		
			(0.009)	(0.009)		
isfake			0.028***	0.028***	0.021*	
			(0.007)	(0.007)	(0.009)	
Tai × isreal				0.018		
				(0.016)		
Tai × isfake					−0.010	
					(0.013)	
engr						0.013
						(0.010)
progexp						0.005
						(0.010)
edulevel						−0.005
						(0.009)
Constant	−0.287***	−0.164***	−0.307***	−0.305***	−0.296***	−0.297***
	(0.023)	(0.017)	(0.025)	(0.024)	(0.023)	(0.023)
Adj. R^2	0.3103	0.3433	0.3147	0.3151	0.3119	0.3114
No. cluster	87	87	87	87	87	87
No. Obs.	2610	2610	2610	2610	2610	2610

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table A.4: WOA and AI-generated Proportion

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.294***	0.266***	0.289***	0.253***
	(0.042)	(0.043)	(0.042)	(0.039)	(0.047)
HMpro	0.0002	0.0005			
	(0.0003)	(0.001)			
Tai × HMpro		−0.001			
		(0.001)			
isreal			0.025	0.061	0.025
			(0.034)	(0.057)	(0.034)
isfake			−0.012	−0.012	−0.032
			(0.024)	(0.024)	(0.037)
Tai × isreal				−0.070	
				(0.060)	
Tai × isfake					0.039
					(0.042)
Constant	0.317***	0.302***	0.322***	0.310***	0.328***
	(0.027)	(0.032)	(0.027)	(0.028)	(0.030)
Adj. R^2	0.0408	0.0407	0.0408	0.0410	0.0406
No. cluster	87	87	87	87	87
No. Obs.	2494	2494	2494	2494	2494

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 116 observations have been excluded where WOA was undefined.

B Additional Experiments: Supplementary Results

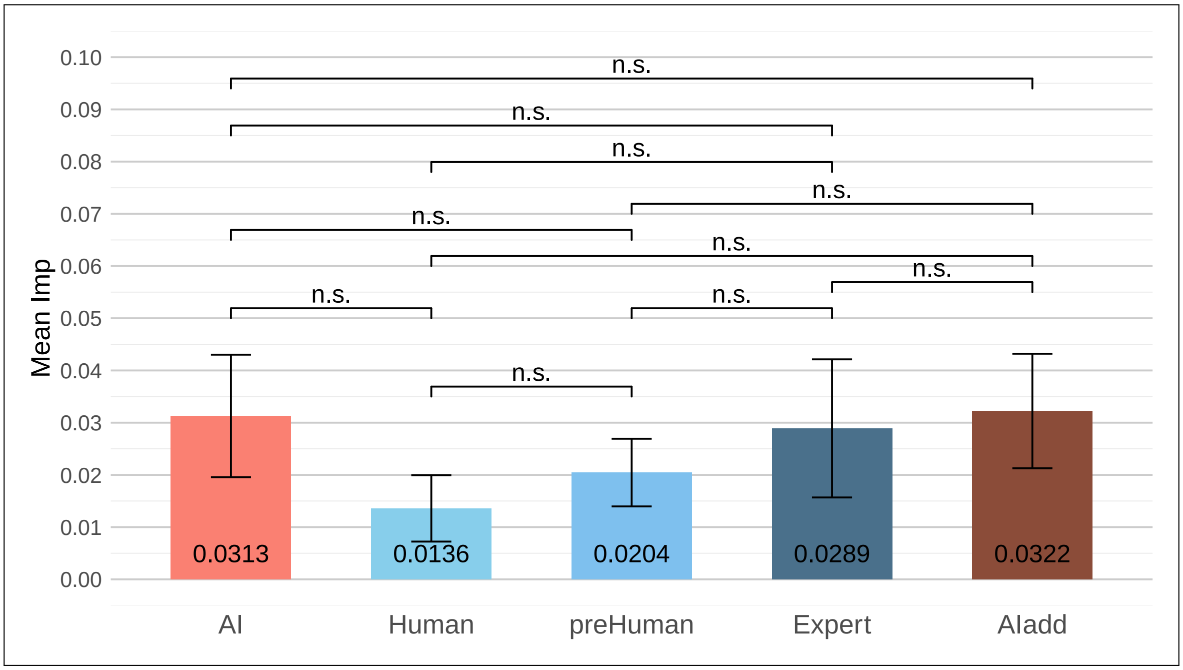


Figure B.1: Mean Imp Across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. Imp is compared across treatments using the Mann–Whitney U test. All reported p-values are adjusted for multiple comparisons using the Holm method.

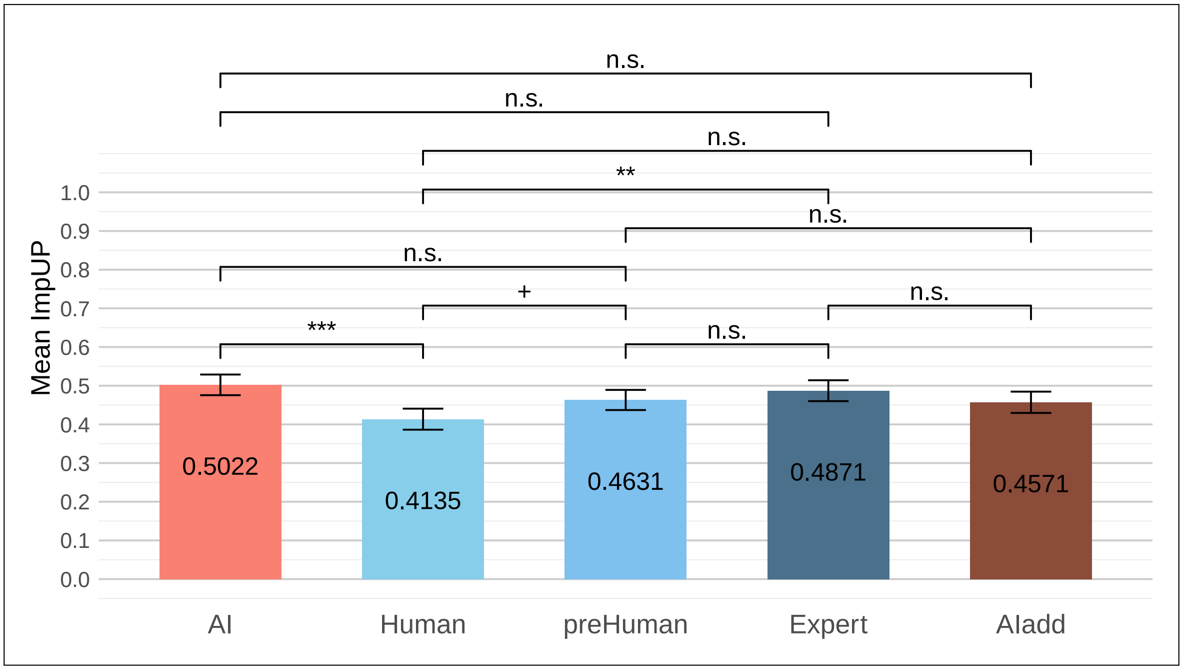


Figure B.2: Mean ImpUP Across Treatments

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. ImpUP is compared across treatments using Fisher's exact tests. All reported p-values are adjusted for multiple comparisons using the Holm method.

Table B.1: Baseline Treatment Effects on Performance (Additional Experiment)

Dep. Var.	$Accu_1$	$Accu_1$	$Accu_1$	$Accu_2$	$Accu_2$	$Accu_2$
	(1)	(2)	(3)	(4)	(5)	(6)
Tai	-0.011	-0.011	-0.014	0.006	0.006	0.004
	(0.014)	(0.014)	(0.013)	(0.009)	(0.009)	(0.009)
Texpert	-0.015	-0.015	-0.014	0.015	0.015	0.015
	(0.015)	(0.015)	(0.015)	(0.010)	(0.010)	(0.010)
TpreHuman	-0.014	-0.014	-0.015	-0.006	-0.006	-0.006
	(0.014)	(0.014)	(0.013)	(0.011)	(0.011)	(0.011)
Taiadd	-0.026^{+}	-0.026^{+}	-0.026^{+}	-0.009	-0.009	-0.009
	(0.015)	(0.015)	(0.015)	(0.011)	(0.011)	(0.011)
HMpro	-0.0001		-0.0001	-0.0003^{**}		-0.0003^{**}
	(0.0001)		(0.0001)	(0.0001)		(0.0001)
isreal		-0.035^{**}			-0.016^{*}	
		(0.011)			(0.008)	
isfake		-0.035^{***}			0.001	
		(0.010)			(0.008)	
engr			-0.023^{*}			-0.011
			(0.012)			(0.008)
progexp			-0.006			-0.001
			(0.011)			(0.008)
edulevel			0.004			0.0004
			(0.010)			(0.008)
Advq				0.546^{***}	0.548^{***}	0.546^{***}
				(0.019)	(0.019)	(0.019)
Constant	0.718^{***}	0.737^{***}	0.740^{***}	0.347^{***}	0.338^{***}	0.355^{***}
	(0.011)	(0.010)	(0.022)	(0.016)	(0.016)	(0.020)
Adj. R^2	0.0008	0.0053	0.0023	0.3811	0.3798	0.3814
No. cluster	220	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table B.2: Interaction Effects Between Advice Source and Advice Quality

Dep. Var.	$Accu_2$					
	(1)	(2)	(3)	(4)	(5)	(6)
Tai	-0.155^{***}	0.006	0.006	0.006	0.005	0.006
	(0.032)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Texpert	0.014	0.050	0.016	0.015	0.015	0.015
	(0.010)	(0.035)	(0.010)	(0.010)	(0.010)	(0.010)
TpreHuman	-0.006	-0.005	0.097^{**}	-0.006	-0.006	-0.006
	(0.011)	(0.011)	(0.031)	(0.011)	(0.011)	(0.011)
Taiadd	-0.009	-0.009	-0.009	-0.032	-0.010	-0.009
	(0.011)	(0.011)	(0.011)	(0.038)	(0.011)	(0.011)
HMpro	-0.0002^{**}	-0.0003^{**}	-0.0003^{**}	-0.0003^{**}	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
isreal					0.075^{***}	0.041^{***}
					(0.019)	(0.010)
isfake					-0.058^{***}	-0.060^{**}
					(0.009)	(0.020)
Advq	0.499^{***}	0.562^{***}	0.569^{***}	0.541^{***}	0.563^{***}	0.542^{***}
	(0.021)	(0.022)	(0.022)	(0.021)	(0.020)	(0.021)
Tai × Advq	0.223^{***}					
	(0.040)					
Texpert × Advq		-0.049				
		(0.044)				
TpreHuman × Advq			-0.143^{***}			
			(0.040)			
Taiadd × Advq				0.032		
				(0.052)		
isreal × Advq					-0.051^{*}	
					(0.023)	
isfake × Advq						0.0002
						(0.022)
Constant	0.380^{***}	0.336^{***}	0.331^{***}	0.351^{***}	0.386^{***}	0.403^{***}
	(0.018)	(0.018)	(0.018)	(0.017)	(0.019)	(0.020)
Adj. R^2	0.3915	0.3817	0.3844	0.3812	0.3851	0.3843
No. cluster	220	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table B.3: Baseline Treatment Effects on Performance Improvement (Additional Experiment)

Dep. Var.	ImpUP	ImpUP	ImpUP	Imp	Imp	Imp
	(1)	(2)	(3)	(4)	(5)	(6)
Tai	0.219*	0.220*	0.229*	0.017*	0.017*	0.018*
	(0.091)	(0.091)	(0.090)	(0.008)	(0.008)	(0.008)
Texpert	0.225*	0.226*	0.214*	0.027***	0.027***	0.026***
	(0.093)	(0.093)	(0.090)	(0.007)	(0.007)	(0.007)
TpreHuman	0.147^{+}	0.147^{+}	0.162^{+}	0.008	0.008	0.009^{+}
	(0.088)	(0.088)	(0.088)	(0.005)	(0.005)	(0.005)
Taiadd	0.086	0.087	0.067	0.017*	0.017*	0.017*
	(0.107)	(0.107)	(0.101)	(0.008)	(0.008)	(0.008)
HMpro	−0.001	0.0002	−0.001	−0.0002***	−0.0002	−0.0002***
	(0.001)	(0.001)	(0.001)	(0.0001)	(0.0001)	(0.0001)
Advq	3.053***	3.056***	3.064***	0.421***	0.423***	0.421***
	(0.109)	(0.108)	(0.108)	(0.018)	(0.019)	(0.018)
isreal		−0.028			0.022**	
		(0.076)			(0.008)	
isfake		0.071			0.026***	
		(0.078)			(0.008)	
engr			0.151*			0.012^{+}
			(0.059)			(0.007)
progexp			0.149*			0.005
			(0.058)			(0.007)
edulevel			−0.081			−0.004
			(0.062)			(0.005)
Constant	−2.438***	−2.500***	−2.715***	−0.278***	−0.298***	−0.292***
	(0.109)	(0.134)	(0.146)	(0.014)	(0.017)	(0.018)
AIC	7431.1	7434.4	7413			
Adj. R^2				0.318	0.3215	0.3184
No. cluster	220	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Models (1)–(3) report Probit regressions, while Models (4)–(6) report OLS regressions. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table B.4: Advice Source and the Likelihood and Magnitude of Improvement

Dep. Var.	ImpUP	ImpUP	ImpUP	ImpUP	Imp	Imp	Imp	Imp
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tai	−0.188	0.219^{*}	0.219^{*}	0.220^{*}	-0.082^{*}	0.017^{*}	0.017^{*}	0.017^{*}
	(0.192)	(0.091)	(0.091)	(0.090)	(0.033)	(0.008)	(0.008)	(0.008)
Texpert	0.225^{*}	0.354^{+}	0.224^{*}	0.225^{*}	0.026^{***}	−0.035	0.027^{***}	0.027^{***}
	(0.091)	(0.198)	(0.093)	(0.092)	(0.007)	(0.029)	(0.008)	(0.007)
TpreHuman	0.146^{+}	0.147^{+}	0.246	0.146^{+}	0.008	0.008	0.138^{***}	0.008
	(0.087)	(0.088)	(0.217)	(0.087)	(0.005)	(0.005)	(0.022)	(0.005)
Taiadd	0.088	0.086	0.086	−0.174	0.018^{*}	0.018^{*}	0.017^{*}	−0.010
	(0.105)	(0.107)	(0.107)	(0.208)	(0.008)	(0.008)	(0.008)	(0.032)
Advq	2.950^{***}	3.103^{***}	3.082^{***}	2.999^{***}	0.393^{***}	0.396^{***}	0.451^{***}	0.416^{***}
	(0.121)	(0.120)	(0.119)	(0.120)	(0.020)	(0.021)	(0.021)	(0.021)
Tai × Advq	0.534^{*}				0.137^{**}			
	(0.259)				(0.047)			
Texpert × Advq		−0.170				0.088^{*}		
		(0.257)				(0.043)		
TpreHuman × Advq			−0.134				-0.181^{***}	
			(0.290)				(0.031)	
Taiadd × Advq				0.339				0.038
				(0.268)				(0.046)
Constant	-2.396^{***}	-2.512^{***}	-2.496^{***}	-2.433^{***}	-0.269^{***}	-0.271^{***}	-0.311^{***}	-0.285^{***}
	(0.109)	(0.109)	(0.109)	(0.109)	(0.015)	(0.015)	(0.015)	(0.015)
AIC	7428.1	7433.5	7434	7432.1				
Adj. R^2					0.3213	0.3186	0.3233	0.316
No. cluster	220	220	220	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Models (1)–(4) report Probit regressions, while Models (5)–(8) report OLS regressions. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table B.5: WOA and AI-generated Proportion (Additional Experiment, *HMpro*)

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.260***	0.266***	0.266***	0.266***
	(0.042)	(0.039)	(0.042)	(0.042)	(0.042)
TpreHuman	−0.004	−0.004	0.020	−0.004	−0.004
	(0.035)	(0.035)	(0.039)	(0.035)	(0.035)
Texpert	0.225***	0.225***	0.225***	0.254***	0.225***
	(0.041)	(0.041)	(0.041)	(0.041)	(0.041)
Taiadd	0.179***	0.179***	0.179***	0.179***	0.169***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.050)
Tai × HMpro		0.0001			
		(0.0003)			
TpreHuman × HMpro			−0.0005		
			(0.0004)		
Texpert × HMpro				-0.001^{+}	
				(0.0003)	
Taiadd × HMpro					0.0002
					(0.0003)
HMpro	−0.0002	−0.0002	−0.0001	−0.0001	−0.0002
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	0.335***	0.337***	0.330***	0.329***	0.337***
	(0.026)	(0.026)	(0.026)	(0.026)	(0.026)
Adj. R^2	0.0396	0.0394	0.0396	0.0397	0.0394
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table B.6: WOA and AI-generated Proportion (Additional Experiment, *isreal*)

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.265***	0.266***	0.266***	0.266***
	(0.042)	(0.039)	(0.042)	(0.042)	(0.042)
TpreHuman	−0.004	−0.004	0.010	−0.004	−0.004
	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)
Texpert	0.225***	0.225***	0.225***	0.240***	0.225***
	(0.041)	(0.041)	(0.041)	(0.040)	(0.041)
Taiadd	0.179***	0.179***	0.179***	0.179***	0.178***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)
Tai × isreal		0.002			
		(0.030)			
TpreHuman × isreal			−0.043		
			(0.032)		
Texpert × isreal				-0.045^{+}	
				(0.025)	
Taiadd × isreal					0.004
					(0.030)
isreal	−0.004	−0.004	0.005	0.005	−0.005
	(0.014)	(0.017)	(0.017)	(0.017)	(0.017)
Constant	0.327***	0.328***	0.324***	0.324***	0.328***
	(0.026)	(0.025)	(0.025)	(0.026)	(0.026)
Adj. R^2	0.0394	0.0392	0.0395	0.0395	0.0392
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table B.7: WOA and AI-generated Proportion (Additional Experiment, $isfake$)

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.268***	0.266***	0.266***	0.266***
	(0.042)	(0.044)	(0.042)	(0.042)	(0.042)
TpreHuman	−0.004	−0.004	−0.020	−0.004	−0.004
	(0.035)	(0.035)	(0.036)	(0.035)	(0.035)
Texpert	0.225***	0.225***	0.225***	0.212***	0.225***
	(0.041)	(0.041)	(0.041)	(0.043)	(0.041)
Taiadd	0.179***	0.179***	0.179***	0.179***	0.189***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.050)
Tai × isfake		−0.008			
		(0.025)			
TpreHuman × isfake			0.047		
			(0.032)		
Texpert × isfake				0.040	
				(0.026)	
Taiadd × isfake					−0.030
					(0.025)
isfake	0.001	0.002	−0.010	−0.007	0.006
	(0.012)	(0.014)	(0.013)	(0.014)	(0.014)
Constant	0.326***	0.325***	0.329***	0.328***	0.324***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Adj. R^2	0.0394	0.0392	0.0395	0.0394	0.0393
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table B.8: WOA and Prior Beliefs

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.286***	0.240***	0.296***	0.286***	0.278***
	(0.041)	(0.051)	(0.041)	(0.041)	(0.041)
Texpert	0.238***	0.236***	0.332***	0.238***	0.237***
	(0.046)	(0.045)	(0.058)	(0.046)	(0.044)
TpreHuman	0.029	0.021	0.048	0.029	0.019
	(0.039)	(0.039)	(0.040)	(0.044)	(0.038)
Taiadd	0.194***	0.191***	0.204***	0.194***	0.119*
	(0.048)	(0.049)	(0.048)	(0.048)	(0.053)
prefAdvSrc	0.086**	0.061^{+}	0.127***	0.086**	0.050^{+}
	(0.029)	(0.033)	(0.032)	(0.033)	(0.029)
freqGPT	−0.001	−0.002	−0.003	−0.001	−0.002
	(0.007)	(0.007)	(0.007)	(0.007)	(0.006)
Tai × prefAdvSrc		0.120^{+}			
		(0.063)			
Texpert × prefAdvSrc			−0.188**		
			(0.070)		
TpreHuman × prefAdvSrc				−0.002	
				(0.058)	
Taiadd × prefAdvSrc					0.173*
					(0.082)
Constant	0.279***	0.293***	0.258***	0.278***	0.301***
	(0.033)	(0.033)	(0.034)	(0.034)	(0.033)
Adj. R^2	0.0442	0.0457	0.0482	0.0441	0.0474
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table B.9: WOA and Advice Quality (Advq)

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.310***	0.266***	0.266***	0.266***
	(0.042)	(0.062)	(0.042)	(0.042)	(0.042)
Texpert	0.229***	0.230***	0.186***	0.229***	0.229***
	(0.041)	(0.041)	(0.053)	(0.041)	(0.041)
TpreHuman	−0.003	−0.003	−0.004	−0.025	−0.003
	(0.035)	(0.035)	(0.035)	(0.059)	(0.035)
Taiadd	0.178***	0.178***	0.179***	0.178***	0.188**
	(0.048)	(0.048)	(0.048)	(0.048)	(0.062)
Advq	0.122***	0.135***	0.103***	0.118***	0.125***
	(0.024)	(0.026)	(0.029)	(0.025)	(0.027)
Tai × Advq		−0.061			
		(0.061)			
Texpert × Advq			0.062		
			(0.051)		
TpreHuman × Advq				0.030	
				(0.075)	
Taiadd × Advq					−0.014
					(0.051)
Constant	0.238***	0.229***	0.252***	0.242***	0.237***
	(0.030)	(0.030)	(0.031)	(0.030)	(0.031)
Adj. R^2	0.0442	0.0422	0.0422	0.0421	0.0421
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table B.10: WOA and Advice Quality (AdvqLag)

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.258^{***}	0.246^{**}	0.258^{***}	0.258^{***}	0.258^{***}
	(0.043)	(0.078)	(0.043)	(0.043)	(0.043)
Texpert	0.218^{***}	0.218^{***}	0.242^{***}	0.218^{***}	0.219^{***}
	(0.042)	(0.042)	(0.054)	(0.042)	(0.042)
TpreHuman	-0.012	-0.012	-0.012	-0.074	-0.012
	(0.037)	(0.037)	(0.036)	(0.058)	(0.036)
Taiadd	0.169^{***}	0.169^{***}	0.169^{***}	0.170^{***}	0.248^{***}
	(0.048)	(0.048)	(0.048)	(0.048)	(0.060)
AdvqLag	0.079^{**}	0.075^{*}	0.090^{*}	0.065^{+}	0.099^{**}
	(0.030)	(0.032)	(0.038)	(0.034)	(0.035)
Tai × AdvqLag		0.017			
		(0.085)			
Texpert × AdvqLag			-0.034		
			(0.061)		
TpreHuman × AdvqLag				0.088	
				(0.072)	
Taiadd × AdvqLag					-0.110^{+}
					(0.056)
Constant	0.277^{***}	0.280^{***}	0.270^{***}	0.287^{***}	0.263^{***}
	(0.032)	(0.031)	(0.034)	(0.033)	(0.033)
Adj. R^2	0.0393	0.0392	0.0392	0.0394	0.0395
No. cluster	220	220	220	220	220
No. Obs.	6,089	6,089	6,089	6,089	6,089

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined. 220 observations have been excluded where *AdvqLag* was undefined.

C Divergence in WOA Across Waves

In **Section** 5.2.3 of the main text, we document a cross-wave inconsistency in relative reliance on AI versus Expert advice. Specifically, participants in the **AI** treatment exhibit a higher WOA than those in the **Expert** treatment, whereas participants in the **AIadd** treatment exhibit a significantly lower WOA than those in the **Expert** treatment. A key complication is that the **AI** and **AIadd** implementations differ along two dimensions simultaneously: the timing of the belief questionnaire (elicited after vs. before the main task) and the calendar time at which the experiment was conducted.

Two mechanisms may contribute to this pattern. First, eliciting beliefs prior to the main task may generate a salience or experimenter-demand effect, prompting participants to infer the purpose of the study and thereby changing how they integrate the subsequently provided advice. Importantly, such an effect may operate through reliance behavior even if the *stated* preference measure itself does not change markedly. Second, cohort differences may arise because participants' perceptions of the relative competence of AI tools plausibly evolve over calendar time as public familiarity with GAI increases.

To shed light on these possibilities, we exploit the fact that the **AI** treatment was conducted in two waves (2023.11 and 2024.10), while **AIadd** was conducted later (2025.10). Figure C.1 plots the mean WOA (solid line) and the mean $prefAdvSrc$ (dashed line) across these waves. Both series exhibit qualitatively similar time patterns, with a peak in 2024.10. However, only WOA displays statistically significant cross-wave differences, whereas differences in $prefAdvSrc$ are imprecisely estimated. In sum, the evidence is consistent with the view that changes in reliance over time may reflect a combination of cohort differences and elicitation-induced salience, though our design does not allow these channels to be cleanly disentangled.

Moreover, additional evidence points toward a cohort-based explanation. Participants' self-reported frequency of using ChatGPT in daily life ($freqGPT$) increases monotonically across waves (2023.11 → 2024.10 → 2025.10), consistent with rising exposure to and familiarity with GAI over calendar time.[1] This secular increase suggests that any evolution in participants' perceptions of AI competence is plausibly driven by broader time trends rather than by the elicitation protocol alone. Accordingly, while elicitation-induced salience may still play a role, we interpret the divergence in WOA primarily as reflecting cohort differences over time.

[1] Mean $freqGPT$: 1.20 (2023.11), 2.12 (2024.10), 3.02 (2025.10). Wilcoxon rank-sum tests: 2023.11 vs. 2024.10, $p = 0.048$; 2024.10 vs. 2025.10, $p = 0.102$ (ties present).

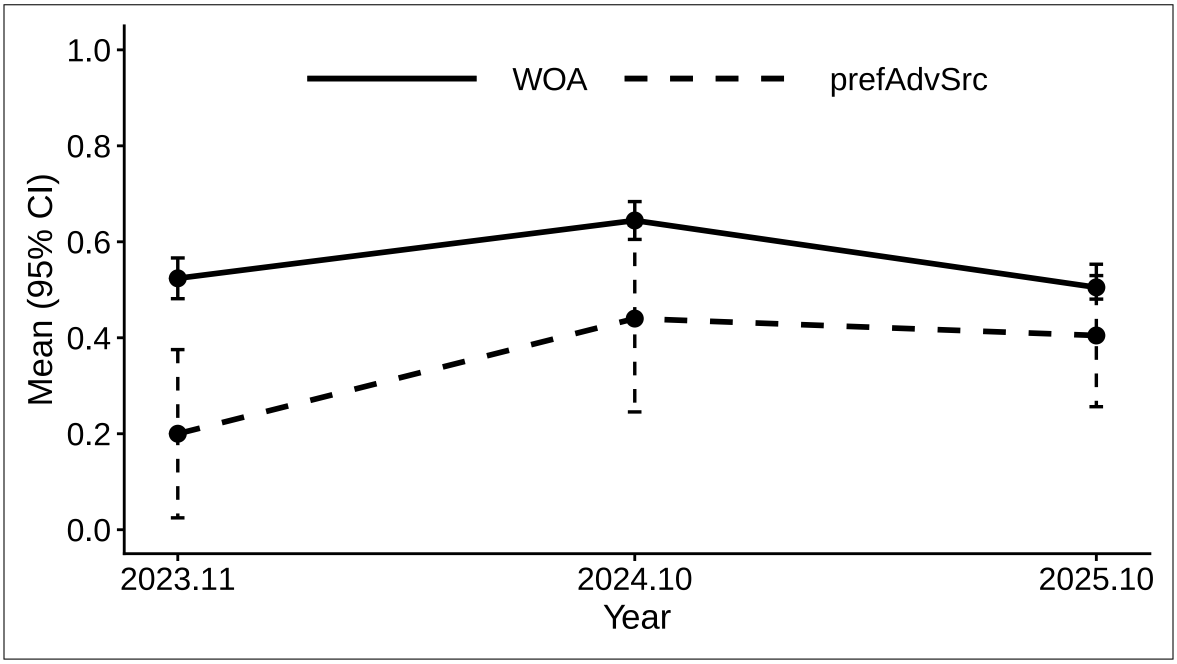


Figure C.1: Time Trends in Reliance and Prior Beliefs (AI Treatments)

Note: The solid line plots the mean WOA and the dashed line plots the mean $prefAdvSrc$ at each wave (2023.11, 2024.10, 2025.10). Mann-Whitney U tests for WOA: 2023.11 vs. 2024.10, $p = 1.26 \times 10^{-7}$; 2024.10 vs. 2025.10, $p < 2.2 \times 10^{-16}$; 2023.11 vs. 2025.10, $p = 0.085$. Fisher exact tests for $prefAdvSrc$: 2023.11 vs. 2024.10, $p = 0.1185$; 2024.10 vs. 2025.10, $p = 0.8029$; 2023.11 vs. 2025.10, $p = 0.1541$. Error bars denote 95% confidence intervals across participants.

D Decomposition of Reliance

D.1 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 identification by at least a threshold (3.5% of the length of the slider they used) amount after receiving the advice. In our analysis, we adopt a more inclusive definition by setting the threshold to zero. Therefore, the activation status of each observation is defined as follows.

$$Acti_r^i = \begin{cases} 1 & \text{if } Final\ Response_r^i \neq Initial\ Response_r^i \\ 0 & \text{if } Final\ Response_r^i = Initial\ Response_r^i \end{cases}$$

Specifically, participant i is considered "activated" in round r if they adjusted their initial response after receiving advice and "not activated" if they maintained their initial response. In our study, 5358 out of the 6600 total samples were activated, resulting in an overall activation rate of 81.2%. The average *Acti* across different treatments are depicted in Figure D.1. All pairwise differences are statistically significant after Holm correction; for clarity, Figure D.1 highlights only the comparisons central to our theoretical arguments.

Results of Probit regressions are shown in Table D.1, where we added the "advice-initial gap (gap between the advice and initial response)" as a new control variable, denoted as $advGap$ and calculated as $|Advice - Initial\ Response|$[2]. Since the quality of the current advice is only realized after the activation decision, we additionally include lagged advice quality ($AdvqLag$) to capture learning from recent advice performance. Consistent with the activation framework, participants are more likely to be activated when the advice is from ChatGPT or from a source they perceive as more effective. Moreover, activation is positively associated with both a larger gap between the advice and the initial identification and higher lagged advice quality.

In sum, the results suggest that activation is jointly shaped by four sets of factors: (i) the source of advice, (ii) prior beliefs about relative advice effectiveness, (iii) the gap between the advice and the initial response, and (iv) recent experience with advice

[2] As Baines et al. (2024) emphasizes, the distance between an advisor's recommendation and a decision-maker's initial (pre-advice) judgment can shape advice utilization and merits more granular measurement. Evidence from human advice settings suggests an inverted-U pattern: decision-makers place more weight on advice when it is neither too close to nor too far from their own initial estimate (Moussaïd et al., 2013; Schultze et al., 2015; Ecken and Pibernik, 2016; Hütter and Ache, 2016). Related evidence in AI advice shows that decision-makers are more likely to follow expert AI advisors when recommendations are closer to their initial judgments (Mesbah et al., 2021).

Table D.1: Probit Regressions of Acti

Dep. Var.	Acti					
	(1)	(2)	(3)	(4)	(5)	(6)
Tai	0.665***	0.665***	0.642***	0.649***	0.649***	0.627***
	(0.147)	(0.147)	(0.152)	(0.148)	(0.148)	(0.153)
Texpert	0.327*	0.327*	0.279^{+}	0.310*	0.310*	0.264
	(0.155)	(0.155)	(0.169)	(0.158)	(0.158)	(0.172)
TpreHuman	0.197	0.197	0.181	0.197	0.197	0.183
	(0.142)	(0.142)	(0.160)	(0.144)	(0.144)	(0.162)
Taiadd	0.269^{+}	0.269^{+}	0.208	0.253	0.253	0.195
	(0.161)	(0.162)	(0.164)	(0.163)	(0.163)	(0.165)
HMpro	0.0001		0.00005	0.0001		0.0001
	(0.0004)		(0.0004)	(0.0004)		(0.0004)
isreal		−0.016			−0.005	
		(0.045)			(0.046)	
isfake		−0.025			−0.012	
		(0.039)			(0.041)	
Advq	0.077	0.076	0.074			
	(0.097)	(0.096)	(0.099)			
AdvqLag				0.188*	0.186*	0.188*
				(0.074)	(0.074)	(0.075)
prefAdvSrc	0.225*	0.225*	0.220*	0.224*	0.224*	0.219*
	(0.095)	(0.095)	(0.093)	(0.096)	(0.096)	(0.095)
advGap	0.019***	0.019***	0.019***	0.018***	0.018***	0.019***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
freqGPT			0.019			0.018
			(0.025)			(0.025)
progexp			0.270**			0.270**
			(0.089)			(0.089)
edulevel			−0.204*			−0.207*
			(0.096)			(0.097)
engr			0.054			0.061
			(0.087)			(0.087)
Constant	0.042	0.060	−0.315	−0.026	−0.014	−0.387*
	(0.145)	(0.144)	(0.193)	(0.138)	(0.140)	(0.183)
AIC	5840	5841	5776	5654	5656	5593
No. cluster	220	220	220	220	220	220
No. Obs.	6600	6600	6600	6380	6380	6380

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 220 observations in models (4)-(6) are excluded because *AdvqLag* is not defined in the first round.

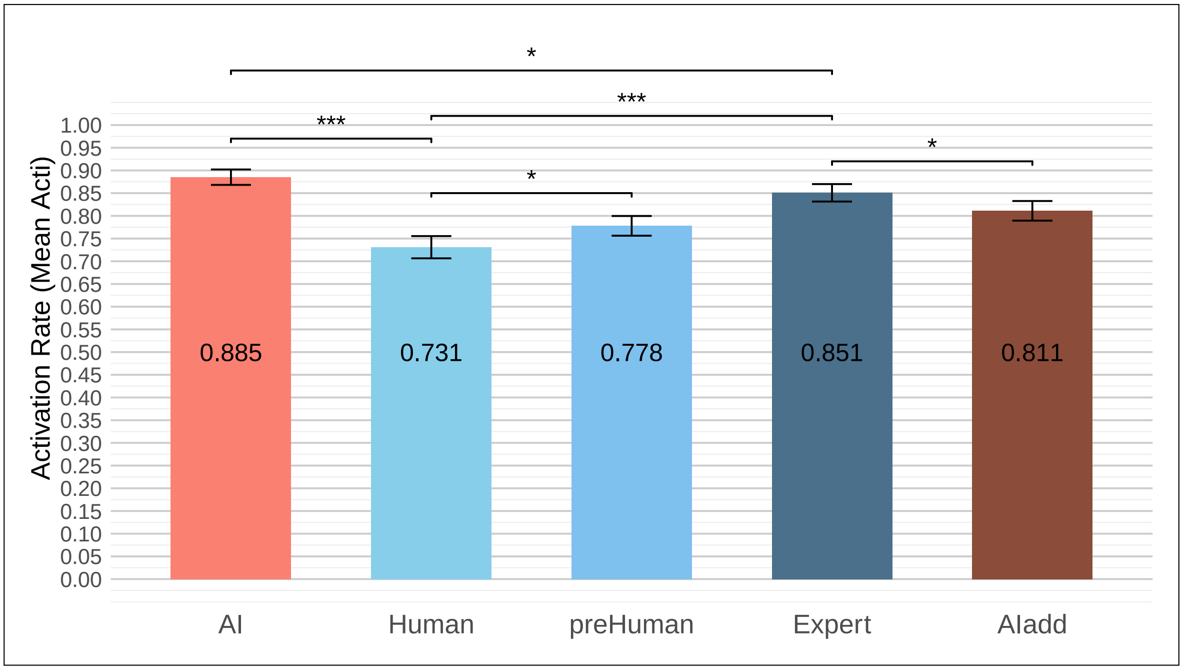


Figure D.1: Mean Acti Across Treatments

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants. Acti is compared across treatments using pairwise proportion tests. All reported p-values are adjusted for multiple comparisons using the Holm method.

quality.

A key result is that contemporaneous advice quality ($Advq$) does not affect activation decisions, whereas lagged advice quality ($AdvqLag$) does. Importantly, advice quality is not directly observable to participants at the time of activation; instead, lagged advice quality captures participants' recent experience with advice performance. In contrast, both contemporaneous and lagged advice quality significantly influence the WOA, as shown in Tables B.9 and B.10. We revisit how advice quality affects reliance using the activation–integration framework in the following subsection.

D.2 Integration Stage

In the second stage of advice taking process (integration), our focus shifts to examining the extent to which participants utilize the advice once they decide to use it. To quantify this extent of utilization, we constructed a continuous variable as follows:

$$Itgr_r^i = \begin{cases} Final\ Response_r^i - Advice_r^i & \text{if } Advice_r^i > Initial\ Response_r^i \\ |Final\ Response_r^i - Advice_r^i| & \text{if } Advice_r^i = Initial\ Response_r^i \\ Advice_r^i - Final\ Response_r^i & \text{if } Advice_r^i < Initial\ Response_r^i \end{cases},$$

which describes the "consistency with the advice" for participant i in round r. 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 Online Appendix I.2. This classification allows us to characterize advice utilization behavior in the integration stage. Table D.2 shows the proportion of these three statuses among the activated sample across different treatments.

Table D.2: Proportion of Utilization Statuses (%)

treatment	Under-utilize	Totally utilize	Over-utilize
AI	76.9	**18.2**	4.9
Human	91	4.3	4.7
preHuman	89.4	5.6	5
Expert	81.6	**13.6**	4.8
AIadd	80.1	**14.3**	5.6
Total	83.5	11.5	5

Participants who received AI or Expert advice are substantially more likely to totally utilize the advice than those receiving advice from human peers. In contrast, the incidence of over-utilization remains remarkably similar across advice sources, suggesting that increased reliance on AI primarily manifests as a higher likelihood of totally following the advice rather than an increased tendency toward extreme over-adjustment.

Here, we apply the Heckman correction approach (Heckman, 1974, 1979) to estimate the two-stage activation–integration process:

First Stage: Activation (Selection)

Activation is modeled using a probit specification. Let

$$Acti_r^i = \mathbb{1}\{z_r^i > 0\},$$

where the latent index z_r^i is given by

$$z_r^i = \alpha_0 + \boldsymbol{\alpha} \cdot \boldsymbol{Treat}^i + \alpha_5\, advGap_r^i + \alpha_6\, prefAdvSrc^i + \alpha_7\, AdvqLag_r^i + \boldsymbol{\alpha_8} \cdot X_1^i + \varepsilon_r^i,$$

where $\varepsilon_r^i \sim \mathcal{N}(0,1)$ and $Acti_r^i = 1$ indicates that participant i adjusted their initial identification after receiving the advice in round r. The selection equation is a probit regression model, estimated on all observations for which $AdvqLag$ is defined (i.e., excluding round 1). The vector $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]$ corresponds to the treatment indicators

$$\boldsymbol{Treat}^i = [Tai^i,\ Texpert^i,\ Taiadd^i,\ TpreHuman^i]'.$$

Second Stage: Integration (Outcome)

Integration is estimated using an OLS specification on the activated subsample (excluding round 1), with selection correction implemented via the inverse Mills ratio (imr)[3] obtained from the first-stage probit model. The outcome equation is specified as:

$$Itgr_r^i = \beta_0 + \boldsymbol{\beta} \cdot \boldsymbol{Treat}^i + \beta_5\, advGap_r^i + \beta_6\, prefAdvSrc^i + \boldsymbol{\beta_7} \cdot X_2^i + \gamma\, imr_r^i + u_r^i,$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, \beta_3, \beta_4]$ and $Itgr_r^i$ measures the extent to which participant i, conditional on activation ($Acti_r^i = 1$), integrates the advice into the final decision in round r. $AdvqLag$ has been excluded as an instrumental variable (exclusion restriction).[4]

[3] The inverse Mills ratio is computed as $imr_r^i = \frac{\phi(\hat{z}_r^i)}{\Phi(\hat{z}_r^i)}$, where $\hat{z}_r^i$ denotes the predicted latent index from the first-stage probit model, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal pdf and cdf. A significant coefficient γ indicates non-random selection into the activated sample, providing evidence of non-random selection.

[4] Identification of the Heckman selection model requires at least one variable that affects selection but does not directly enter the outcome equation (an exclusion restriction). In our specification, $AdvqLag$ is included in the selection (activation) equation but excluded from the outcome equation. Empirically, $AdvqLag$ significantly predicts activation (see Table D.1) but is not significantly associated with uncorrected integration in the OLS specification (see Table D.3). While prior studies do not provide strong theoretical guidance on whether perceived lagged advice quality directly affects the

X_1 and X_2 denote additional control variables included in the selection and outcome equations, respectively.

We applied this method to our analysis, considering that there may be threshold levels for $\boldsymbol{Treat}$, $advGap$, $prefAdvSrc$ and $AdvqLag$. If these variables do not reach certain thresholds, a participant might not alter their initial identification after receiving advice (i.e., not activated). Moreover, by employing this method in the first stage, we can include samples that were not activated. This approach allows us to exploit the full sample and correct for potential selection bias inherent in WOA-based analyses.

Table D.3 reports the results from the Heckman selection model. The coefficient on the inverse Mills ratio (imr) is positive and statistically significant, indicating that selection into *advice activation is non-random and that correcting for selection is empirically relevant.*

Consistent with the activation–integration framework, the advice source exerts a strong influence on the activation stage: participants are substantially more likely to take AI or Expert advice. Conditional on activation, AI or Expert's advice also leads to a greater degree of integration relative to human-peer advice, indicating that the source continues to shape advice taking behavior in both stages of the process.

Participants' prior beliefs about which source performs better ($prefAdvSrc$) likewise influence both stages. Individuals who believe the offered source is more effective in the deepfake-detection task are not only more likely to be activated but also integrate the advice more strongly. This suggests that prior beliefs act as a persistent filter through which advice is interpreted and weighted.

The distance between the advice and the initial response ($advGap$) shows a distinct two-stage pattern: it increases the likelihood of activation but reduces the degree of integration. A larger discrepancy appears to prompt participants to consider taking the advice—perhaps because the difference signals a potentially meaningful correction—yet once activated, the same discrepancy induces caution and limits how far participants are willing to adjust. This pattern aligns with the well-documented inverted-U relationship in advice taking, where decision-makers place the most weight on advice that is neither too close to nor too far from their initial estimate (Moussaïd et al., 2013; Schultze et al., 2015; Ecken and Pibernik, 2016; Hütter and Ache, 2016). Very large gaps may draw attention but ultimately discourage full or over incorporation of the advice.

Overall, the decomposition shows that advice taking follows a two-step process. Activation is shaped mainly by advice source, prior beliefs, the advice–initial gap, and

extent of advice integration conditional on activation, we interpret the Heckman correction as a structural framework distinguishing between activation and integration stages, rather than as providing a strong causal identification of selection bias.

Table D.3: Uncorrected OLS and Heckman Selection Model

	Uncorrected OLS	Heckman (2-step)	
	Integration	Activation	Integration
Dep. Var.	Itgr	Acti	Itgr
Tai	7.724^{***}	0.649^{***}	16.124^{***}
	(0.988)	(0.148)	(2.781)
Texpert	4.790^{***}	0.310^{***}	9.023^{***}
	(1.388)	(0.158)	(1.969)
TpreHuman	-0.800	0.197	2.214^{+}
	(0.975)	(0.144)	(1.275)
Taiadd	4.565^{***}	0.253	8.414^{***}
	(1.248)	(0.163)	(1.641)
AdvqLag	0.432	0.188^{*}	
	(0.737)	(0.074)	
HMpro	-0.006^{+}	0.0001	-0.006^{+}
	(0.003)	(0.0004)	(0.003)
advGap	-0.500^{***}	0.018^{***}	-0.299^{***}
	(0.025)	(0.003)	(0.068)
prefAdvSrc	1.109	0.224^{*}	3.945^{**}
	(0.890)	(0.096)	(1.304)
freqGPT	0.279		0.273
	(0.270)		(0.212)
engr	2.429^{*}		2.361^{*}
	(0.974)		(0.981)
progexp	2.122^{*}		2.119^{*}
	(0.980)		(0.989)
edulevel	-0.583		-0.451
	(0.935)		(0.931)
imr			35.658^{***}
			(9.351)
Constant	-8.200^{***}	-0.026	-29.874^{***}
	(1.995)	(0.138)	(6.764)
No. cluster	220	220	220
No. Obs.	5176	6380	5176

Note: $^{+}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. The Heckman selection model is estimated using the two-step (Heckit) procedure. Selection correction in the outcome equation is implemented via the inverse Mills ratio (*imr*), constructed from the first-step probit selection equation. 220 observations are excluded because *AdvqLag* is not defined in the first round. 1242 inactivated observations are excluded in the two Integration specifications.

lagged advice quality, whereas integration depends on prior beliefs, the gap, and the advice source. The gap operates in opposite directions across the two stages—encouraging activation but limiting integration—while lagged advice quality matters only for activation. The significant selection term further indicates that activation is non-random, so treatment differences in reliance arise from source-driven activation combined with belief- and gap-driven integration.

E Other Robustness Checks

E.1 Demographics

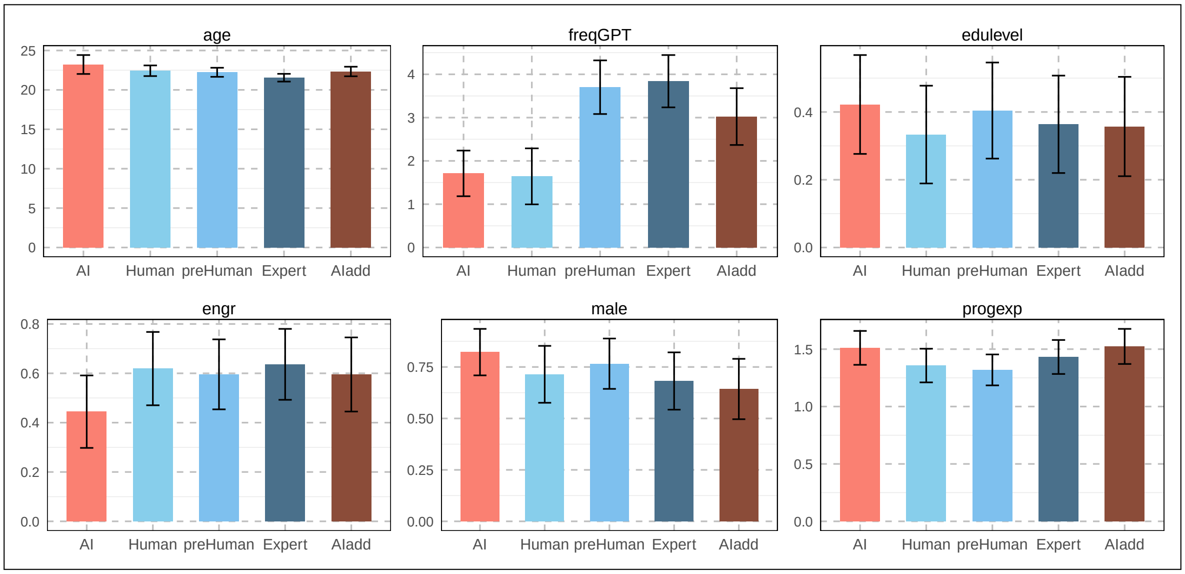


Figure E.1: Demographic Comparisons

Note: Bars report the mean of participants' reported values. Error bars denote 95% confidence intervals across participants.

Figure E.1 presents comparisons of demographic characteristics across treatments. Table E.1 reports OLS regressions examining heterogeneity along demographic dimensions. Overall, demographic variables account for little variation in either WOA or performance outcomes, and controlling for these characteristics leaves the estimated treatment effects largely unchanged.

Table E.1: Demographic Controls Across Outcomes

Dep. Var.	WOA	Accu1	Accu2	Imp
	(1)	(2)	(3)	(4)
Tai	0.275^{***}	-0.015	0.003	0.018^{*}
	(0.039)	(0.013)	(0.009)	(0.008)
Texpert	0.215^{***}	-0.009	0.007	0.015^{+}
	(0.044)	(0.015)	(0.012)	(0.008)
TpreHuman	0.002	-0.005	0.001	0.006
	(0.035)	(0.014)	(0.012)	(0.006)
Taiadd	0.164^{***}	-0.018	-0.001	0.017^{*}
	(0.045)	(0.015)	(0.012)	(0.009)
age	0.004	-0.003^{+}	0.0001	0.003^{**}
	(0.005)	(0.002)	(0.001)	(0.001)
freqGPT	0.003	-0.006^{**}	-0.005^{**}	0.001
	(0.007)	(0.002)	(0.002)	(0.001)
progexp	0.087^{**}	-0.008	-0.001	0.007
	(0.032)	(0.011)	(0.008)	(0.007)
male	-0.012	0.002	-0.0003	-0.002
	(0.032)	(0.010)	(0.007)	(0.006)
edulevel	-0.069^{*}	0.020^{+}	0.006	-0.014^{+}
	(0.034)	(0.011)	(0.008)	(0.007)
engr	0.109^{***}	-0.023^{*}	-0.012	0.011
	(0.030)	(0.011)	(0.009)	(0.007)
Constant	0.067	0.799^{***}	0.736^{***}	-0.063^{*}
	(0.120)	(0.041)	(0.029)	(0.028)
Adj. R^2	0.0481	0.0051	0.0016	0.0018
No. cluster	220	220	220	220
No. Obs.	6309	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

E.2 Decision Time

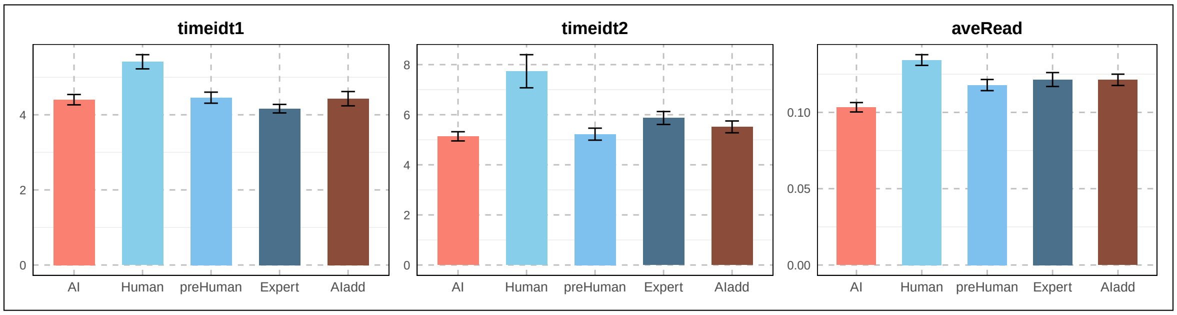


Figure E.2: Decision Time Comparisons

Note: Error bars denote 95% confidence intervals across participants.

We record participants' decision time at different stages of the task, including the time spent on the first identification (*timeidt*1), the time spent on the second identification (*timeidt*2), and the time spent reading the news article. Because news articles differ in length, reading time is normalized as the average time a participant spends per character of the article (*aveRead*).

On average, participants spend 44.68 seconds reading each article (corresponding to 0.12 seconds per character), 4.56 seconds on the first identification, and 5.87 seconds on the second identification. Figure E.2 compares decision time across treatments. In the main experiment, participants in the AI treatment spend significantly less time on reading and identification than those in the Human treatment, suggesting that participants receiving human advice are relatively more time-intensive in their decision-making process (Mann–Whitney U tests with Holm-adjusted $p < 0.001$).

Table E.2 reports OLS regressions examining heterogeneity in reliance and performance outcomes with respect to decision time. In sum, decision-time variables explain little variation in either reliance or performance, and their inclusion does not materially affect the estimated treatment effects.

Table E.2: Time Controls Across Outcomes

Dep. Var.	WOA	Accu1	Accu2	Imp
	(1)	(2)	(3)	(4)
Tai	0.263^{***}	-0.007	0.009	0.016^{*}
	(0.043)	(0.014)	(0.010)	(0.008)
Texpert	0.224^{***}	-0.017	-0.001	0.016^{+}
	(0.042)	(0.014)	(0.011)	(0.008)
TpreHuman	-0.005	-0.010	-0.004	0.006
	(0.037)	(0.014)	(0.012)	(0.005)
Taiadd	0.178^{***}	-0.023	-0.003	0.019^{*}
	(0.048)	(0.015)	(0.012)	(0.008)
timeidt1	-0.001	0.001	0.002^{+}	0.001
	(0.003)	(0.001)	(0.001)	(0.001)
timeidt2	0.001	0.0005	0.0004	-0.0001
	(0.001)	(0.0004)	(0.0004)	(0.0003)
aveRead	-0.138	0.102^{+}	-0.002	-0.104^{*}
	(0.164)	(0.058)	(0.054)	(0.041)
roundnum	0.0003	0.003^{***}	0.004^{***}	0.001^{*}
	(0.001)	(0.0003)	(0.0003)	(0.0003)
Constant	0.337^{***}	0.635^{***}	0.646^{***}	0.012
	(0.052)	(0.015)	(0.014)	(0.009)
Adj. R^2	0.0395	0.0005	0.0004	0.0024
No. cluster	220	220	220	220
No. Obs.	6309	6600	6600	6600

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. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

E.3 News Categories

The 30 news articles are categorized based on their content and primary topic, as reported in Table E.3. Among them, 3 articles are classified as POL, 2 as BUS, 11 as SCI, and 14 as CUL.

Because some categories (in particular BUS and POL) contain relatively few observations, for robustness analyses the categories are aggregated into **HARD** (POL, BUS, and SCI), which correspond to more fact-oriented and policy-relevant domains, and **SOFT** (CUL), which covers lifestyle, cultural, sports, and entertainment topics.

We examine whether reliance and performance outcomes vary with news categories. Table E.3 reports OLS estimates of WOA with category indicators and treatment-by-category interactions. Across specifications, neither $HARDnews$ nor the interaction terms are statistically significant, suggesting no systematic heterogeneity in WOA between **HARD** and **SOFT** news.

We also test for performance heterogeneity by estimating OLS models for $Accu_1$, $Accu_2$, and Imp, and probit models for $ImpUP$ (Tables E.5–E.8). A consistent pattern emerges. First, initial accuracy is lower for **HARD** news: $HARDnews$ is significantly negative in the $Accu_1$ regressions (Tables E.5), indicating that **HARD** items are more difficult in the baseline (no-advice) stage. Second, AI advice appears particularly helpful for **HARD** news: the interaction $Tai \times HARDnews$ is positive for $Accu_1$ and strongly positive for $Accu_2$ (Tables E.5 and Tables E.6), and it is also positive for performance improvement (Imp; Tables E.7). In the AIadd treatment, $Taiadd \times HARDnews$ is strongly positive for $Accu_2$, Imp, and $ImpUP$ (Tables E.6–Tables E.8), indicating larger gains on **HARD** news at later stages.

Overall, while reliance does not differ systematically by news category, participants exhibit lower accuracy when detecting **HARD** deepfake news compared to detecting **SOFT** deepfake news, and AI-based detection can partially mitigate this disadvantage.

Table E.3: News Category with Short Descriptions

Round	Category	Short Description
1	POL	International dispute over ivory trade.
2	SCI	Cesium soil contamination linked to Chernobyl.
3	SCI	Hurricane crossing the dateline and becoming a typhoon.
4	CUL	Celebrity-related marriage and proposal anecdote.
5	CUL	Biography of a Living National Treasure in Noh theater.
6	CUL	Proposal for a new professional baseball development league.
7	CUL	Death of a former Japanese football association leader.
8	CUL	Rules and participation conditions for a horse-racing event.
9	CUL	Career overview of a television host and entertainer.
10	BUS	Business closure due to financial difficulties.
11	CUL	Corporate amateur baseball championship final.
12	SCI	Large-scale gas explosion accident in San Francisco.
13	POL	Diplomatic dispute concerning UN peacekeeping participation.
14	CUL	Career history of a racehorse and its retirement.
15	CUL	High school baseball regional championship result.
16	SCI	Census statistics on population aging in Japan.
17	SCI	Regulatory standards for certified consumer electronics.
18	BUS	Government debate over consumption tax policy.
19	SCI	Earthquake and subsequent emergency response.
20	CUL	Divorce of well-known entertainment figures.
21	SCI	Death of a giant panda at Ueno Zoo.
22	SCI	Chemical explosion incident on a subway train.
23	CUL	Football league admission and promotion decisions.
24	SCI	Environmental dispute over land reclamation project.
25	SCI	Transition from analog to digital broadcasting.
26	CUL	Death of a former sumo wrestler and stablemaster.
27	CUL	Criminal case involving the death of a child.
28	CUL	Death of a Hollywood actress.
29	SCI	Scientific claims related to extraterrestrial life.
30	POL	Political controversy over prime ministerial responsibility.

Notes: News articles are identified by round number and categorized based on their primary topic. POL = Politics / Public affairs; BUS = Economy / Business; SCI = Science / Tech / Health / Environment; CUL = Society / Lifestyle / Culture / Sports / Entertainment.

Table E.4: WOA and News Category

Dep. Var.	WOA				
	(1)	(2)	(3)	(4)	(5)
Tai	0.266***	0.273***	0.266***	0.266***	0.266***
	(0.042)	(0.048)	(0.042)	(0.042)	(0.042)
TpreHuman	−0.004	−0.004	−0.023	−0.004	−0.004
	(0.035)	(0.035)	(0.042)	(0.035)	(0.035)
Texpert	0.225***	0.225***	0.225***	0.218***	0.225***
	(0.041)	(0.041)	(0.041)	(0.045)	(0.041)
Taiadd	0.179***	0.179***	0.179***	0.179***	0.165***
	(0.048)	(0.048)	(0.048)	(0.048)	(0.049)
Tai × HARDnews		−0.014			
		(0.036)			
TpreHuman × HARDnews			0.036		
			(0.037)		
Texpert × HARDnews				0.014	
				(0.032)	
Taiadd × HARDnews					0.026
					(0.028)
HARDnews	0.015	0.018	0.007	0.013	0.010
	(0.015)	(0.017)	(0.017)	(0.018)	(0.018)
Constant	0.318***	0.316***	0.322***	0.319***	0.321***
	(0.030)	(0.031)	(0.031)	(0.031)	(0.031)
Adj. R^2	0.0396	0.0394	0.0396	0.0394	0.0395
No. cluster	220	220	220	220	220
No. Obs.	6309	6309	6309	6309	6309

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *HARDnews* is an indicator equal to 1 for participant who faces with a piece of **HARD** news. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level. 291 observations have been excluded where WOA was undefined.

Table E.5: Performance ($Accu_1$) and News Category

Dep. Var.	$Accu_1$				
	(1)	(2)	(3)	(4)	(5)
Tai	−0.011	-0.026^{+}	−0.011	−0.011	−0.011
	(0.014)	(0.015)	(0.014)	(0.014)	(0.014)
TpreHuman	−0.014	−0.014	−0.011	−0.014	−0.014
	(0.014)	(0.014)	(0.017)	(0.014)	(0.014)
Texpert	−0.015	−0.015	−0.015	−0.003	−0.015
	(0.015)	(0.015)	(0.015)	(0.017)	(0.015)
Taiadd	-0.026^{+}	-0.026^{+}	-0.026^{+}	-0.026^{+}	−0.025
	(0.015)	(0.015)	(0.015)	(0.015)	(0.018)
Tai × HARDnews		0.029^{*}			
		(0.012)			
TpreHuman × HARDnews			−0.005		
			(0.013)		
Texpert × HARDnews				-0.022^{+}	
				(0.012)	
Taiadd × HARDnews					−0.001
					(0.014)
HARDnews	-0.015^{**}	-0.021^{***}	-0.013^{*}	-0.010^{+}	-0.014^{**}
	(0.005)	(0.006)	(0.006)	(0.006)	(0.005)
Constant	0.721^{***}	0.724^{***}	0.720^{***}	0.719^{***}	0.721^{***}
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Adj. R^2	0.0014	0.0019	0.0013	0.0016	0.0013
No. cluster	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *HARDnews* is an indicator equal to one if a participant is exposed to **HARD** news. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table E.6: Performance ($Accu_2$) and News Category

Dep. Var.	$Accu_2$				
	(1)	(2)	(3)	(4)	(5)
Tai	0.007	-0.023^{+}	0.007	0.007	0.007
	(0.009)	(0.012)	(0.009)	(0.009)	(0.009)
TpreHuman	−0.007	−0.007	0.010	−0.007	−0.007
	(0.012)	(0.012)	(0.014)	(0.012)	(0.012)
Texpert	0.0004	0.0004	0.0004	0.016	0.0004
	(0.011)	(0.011)	(0.011)	(0.014)	(0.011)
Taiadd	−0.008	−0.008	−0.008	−0.008	-0.028^{+}
	(0.012)	(0.012)	(0.012)	(0.012)	(0.015)
Tai × HARDnews		0.056^{***}			
		(0.011)			
TpreHuman × HARDnews			-0.033^{**}		
			(0.012)		
Texpert × HARDnews				-0.029^{*}	
				(0.013)	
Taiadd × HARDnews					0.038^{***}
					(0.011)
HARDnews	0.014^{**}	0.002	0.021^{***}	0.019^{***}	0.006
	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
Constant	0.720^{***}	0.726^{***}	0.716^{***}	0.716^{***}	0.724^{***}
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Adj. R^2	0.0008	0.0032	0.0015	0.0013	0.0017
No. cluster	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *HARDnews* is an indicator equal to one if a participant is exposed to **HARD** news. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table E.7: Performance Improvement (Imp) and News Category

Dep. Var.	Imp				
	(1)	(2)	(3)	(4)	(5)
Tai	0.018*	0.003	0.018*	0.018*	0.018*
	(0.008)	(0.009)	(0.008)	(0.008)	(0.008)
TpreHuman	0.007	0.007	0.022***	0.007	0.007
	(0.005)	(0.005)	(0.006)	(0.005)	(0.005)
Texpert	0.015^{+}	0.015^{+}	0.015^{+}	0.019^{+}	0.015^{+}
	(0.008)	(0.008)	(0.008)	(0.010)	(0.008)
Taiadd	0.019*	0.019*	0.019*	0.019*	−0.003
	(0.008)	(0.008)	(0.008)	(0.008)	(0.010)
Tai × HARDnews		0.027**			
		(0.010)			
TpreHuman × HARDnews			−0.028***		
			(0.007)		
Texpert × HARDnews				−0.008	
				(0.012)	
Taiadd × HARDnews					0.040***
					(0.012)
HARDnews	0.028***	0.023***	0.034***	0.030***	0.021***
	(0.004)	(0.005)	(0.005)	(0.004)	(0.004)
Constant	−0.001	0.002	−0.005	−0.002	0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Adj. R^2	0.0063	0.0070	0.0070	0.0062	0.0078
No. cluster	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *HARDnews* is an indicator equal to one if a participant is exposed to **HARD** news. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

Table E.8: Performance Improvement (ImpUP) and News Category

Dep. Var.	ImpUP				
	(1)	(2)	(3)	(4)	(5)
Tai	0.224^{**}	0.167^{*}	0.224^{**}	0.224^{**}	0.224^{**}
	(0.078)	(0.085)	(0.078)	(0.078)	(0.078)
TpreHuman	0.126^{+}	0.126^{+}	0.160^{+}	0.126^{+}	0.126^{+}
	(0.076)	(0.076)	(0.084)	(0.076)	(0.076)
Texpert	0.186^{*}	0.186^{*}	0.186^{*}	0.246^{**}	0.186^{*}
	(0.078)	(0.078)	(0.078)	(0.089)	(0.078)
Taiadd	0.111	0.111	0.111	0.111	0.021
	(0.090)	(0.090)	(0.090)	(0.090)	(0.095)
Tai × HARDnews		0.107			
		(0.066)			
TpreHuman × HARDnews			-0.063		
			(0.084)		
Texpert × HARDnews				-0.111	
				(0.069)	
Taiadd × HARDnews					0.168^{*}
					(0.069)
HARDnews	0.105^{***}	0.083^{*}	0.118^{***}	0.127^{***}	0.073^{*}
	(0.030)	(0.034)	(0.031)	(0.034)	(0.033)
Constant	-0.275^{***}	-0.263^{***}	-0.282^{***}	-0.287^{***}	-0.257^{***}
	(0.063)	(0.063)	(0.063)	(0.063)	(0.063)
AIC	9094.4	9094.5	9095.7	9094.3	9091.9
No. cluster	220	220	220	220	220
No. Obs.	6600	6600	6600	6600	6600

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *HARDnews* is an indicator equal to one if a participant is exposed to **HARD** news. Numbers in parentheses represent standard errors, which are corrected for within-subject clustering at the participant level.

E.4 Learning Effects

We first examine learning effects using a regression-based approach by including round-level time controls in the outcome regressions. Table E.2 reports estimates from regressions of reliance and performance outcomes on treatment indicators and round-related controls. The coefficient on *roundnum* is small and statistically insignificant for WOA, indicating no systematic learning effect in advice reliance over 30 rounds. In contrast, *roundnum* is positive and statistically significant for all three performance measures, suggesting that participants' detection accuracy and overall performance improve as they gain experience with the task.

To further assess learning dynamics, we compare participant-level outcomes across early (rounds 1–10), middle (rounds 11–20), and late (rounds 21–30) blocks within each treatment. For each outcome, bars report participant-level means and error bars denote 95% confidence intervals across participants. Learning within treatments is evaluated by comparing early and late blocks using paired Wilcoxon signed-rank tests. The results are presented in Figures E.3–E.6.

The block-based evidence is consistent with the regression results in Table E.2. We find no learning effects in advice reliance: WOA does not differ significantly between early and late rounds in any treatment. In contrast, strong learning effects emerge in performance. Across all treatments, participants exhibit higher initial accuracy in the final ten rounds than in the first ten rounds. However, learning in performance improvement is treatment-specific: only participants receiving AI-based advice (**AI** and **AIadd**) show a statistically significant increase in performance improvement from early to late rounds, whereas no such pattern is observed in the human or expert advice conditions.

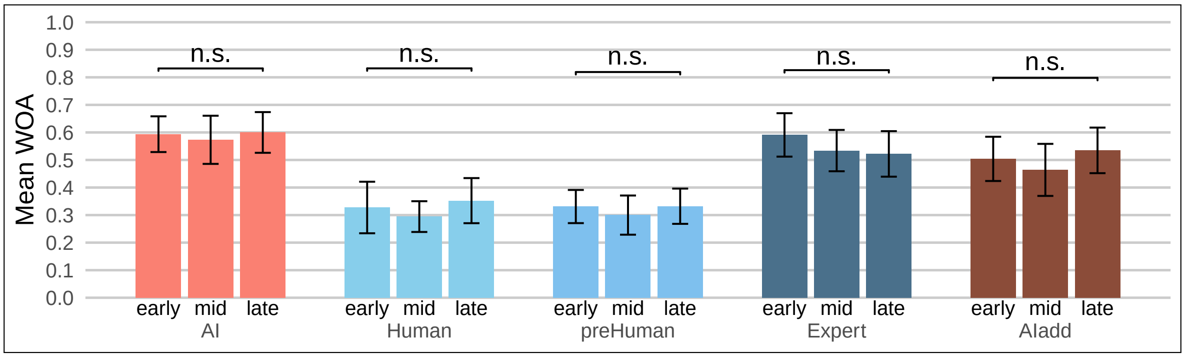


Figure E.3: Reliance Across Early, Middle, and Late Rounds

Note: "n.s." indicates $p \geq 0.1$. No statistically significant differences in reliance are detected between early and late rounds within any treatment.

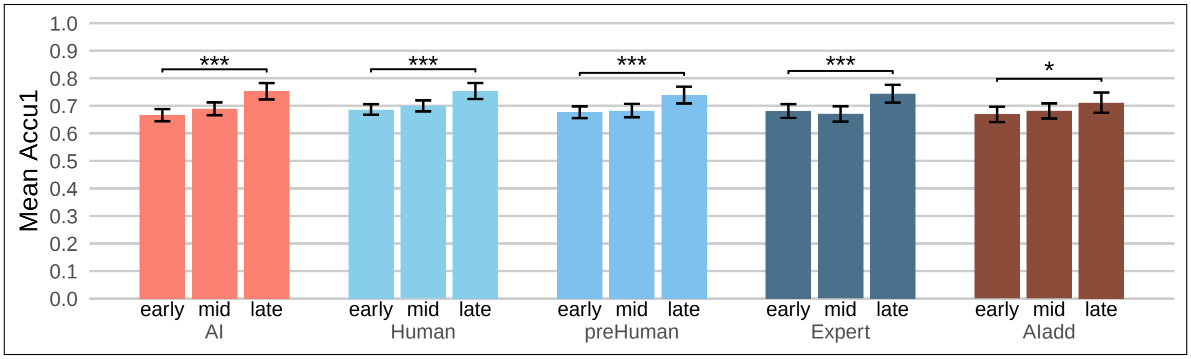


Figure E.4: Initial Accuracy Across Early, Middle, and Late Rounds

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Initial Accuracy increases significantly from early to late rounds within all treatments.

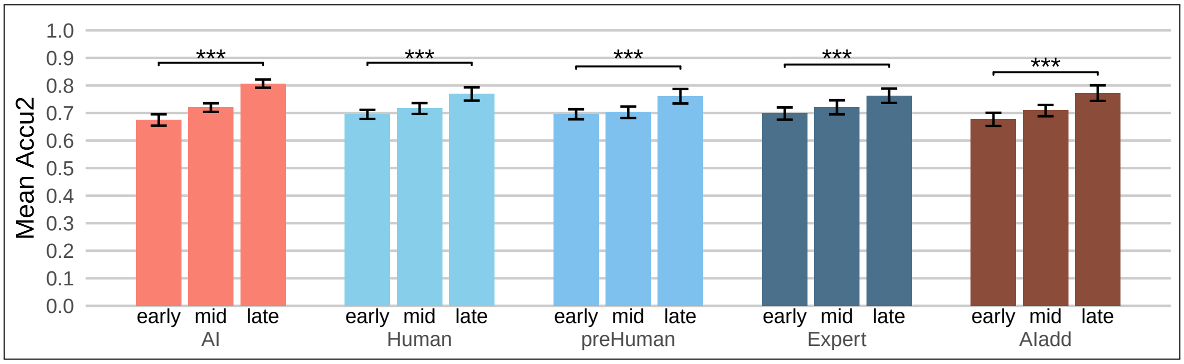


Figure E.5: Final Accuracy Across Early, Middle, and Late Rounds

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Final Accuracy increases significantly from early to late rounds within all treatments.

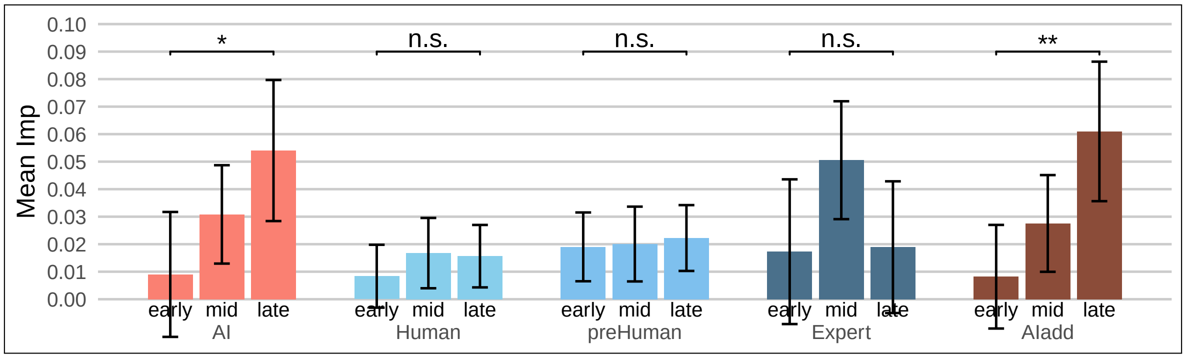


Figure E.6: Performance improvement Across Early, Middle, and Late Rounds

Note: $^+$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. "n.s." indicates $p \geq 0.1$. Performance improvement from early to late rounds is statistically significant in the AI and AIadd treatments, but not in the Human, preHuman, or Expert treatments.

F Experiment Instruction

Treatments **Human, AI** only:

Welcome

- Thank you for participating in this experiment. By taking part in and completing this experiment, we will pay you 500 yen as a participation fee.
- In addition to the 500 yen participation fee, you can earn extra rewards in the decision-making task you will do now.
- During the experiment, please turn off your mobile phone and focus on the experiment. If you have any questions, please ask the experimenter.
- In today's experiment, you will first complete the main decision-making task, and finally answer some questions. Your earnings during the experiment will be paid in private.

Treatments **preHuman, Expert, AIadd** only:

Welcome

- Thank you for participating in this experiment. By taking part in and completing this experiment, we will pay you 500 yen as a participation fee.
- In addition to the 500 yen participation fee, you can earn extra rewards in the decision-making task you will do now.
- During the experiment, please turn off your mobile phone and focus on the experiment. If you have any questions, please ask the experimenter.
- In today's experiment, you will first answer some questions, then complete the main decision-making task. Your earnings during the experiment will be paid in private.

The Main Task

- There are 30 rounds in this experiment.
- In each round, a piece of news will be displayed. Your main task is to identify

the authenticity of that news. Your additional earnings will vary depending on the accuracy of your identifications.

Treatment **Human** only:

The Main Task

- In each round, you will make two identifications about the same news. Please use the slider and report your identification as to its authenticity as an integer from 0 to 100, with 0 representing totally fake news and 100 representing totally real news.
- After your first identification, a number representing **the identification of another randomly selected participant from today's experiment** will be displayed. Based on this, please adjust your first identification and make a second identification. If you think no adjustment is necessary, please report the same result as your first identification.

Treatments **AI, AIadd** only:

The Main Task

- In each round, you will make two identifications about the same news. Please use the slider and report your identification as to its authenticity as an integer from 0 to 100, with 0 representing totally fake news and 100 representing totally real news.
- After your first identification, a number representing **the identification of the AI tool – ChatGPT** will be displayed. Based on this, please adjust your first identification and make a second identification. If you think no adjustment is necessary, please report the same result as your first identifications.

Treatment **preHuman** only:

The Main Task

- In each round, you will make two identifications about the same news. Please use the slider and report your identification as to its authenticity as an integer from 0 to 100, with 0 representing totally fake news and 100 representing totally real news.

- After your first identification, **the number corresponding to the first identification made by one randomly selected participant from a previous (last year's) session of this experiment (42 Osaka University students)** for the same news article will be displayed. Based on this, please adjust your first identification and make a second identification. If you think no adjustment is necessary, please report the same result as your first identifications.

Treatment **Expert** only:

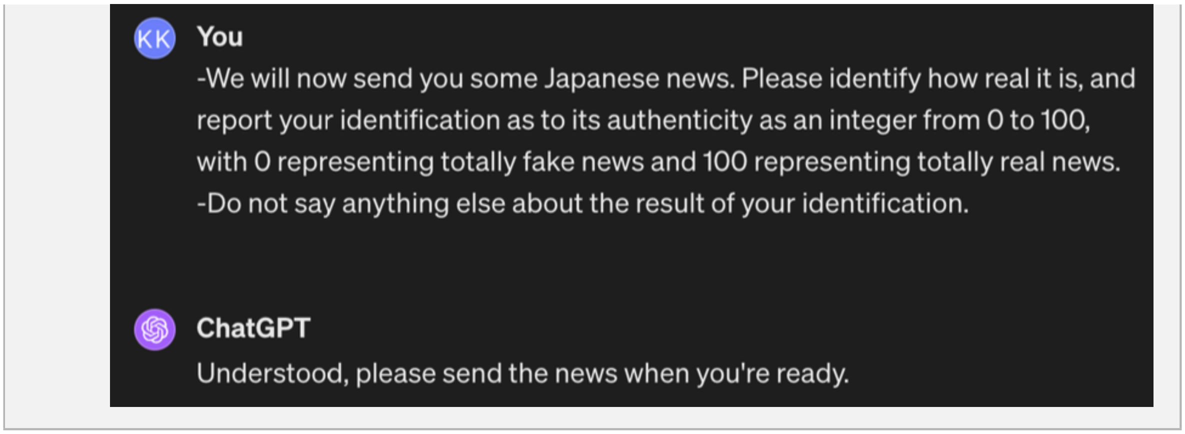
The Main Task

- In each round, you will make two identifications about the same news. Please use the slider and report your identification as to its authenticity as an integer from 0 to 100, with 0 representing totally fake news and 100 representing totally real news.
- After your first identification, you will be shown **the identification value made for the same news article by one randomly selected expert from a pool of 11 linguistics experts (university professors, associate professors, and lecturers specializing in linguistics)**. Based on this, please adjust your first identification and make a second identification. If you think no adjustment is necessary, please report the same result as your first identifications.

Treatments **AI, AIadd** only:

ChatGPT's Identification

- All the news used in today's experiment had already been identified for their 'authenticity' by the AI tool–ChatGPT using prompts before the experiment.
- For all the news, ChatGPT was asked to make a identification 24 times under the same conditions. In each round of the main task, the ChatGPT's identification you will see is randomly selected from these 24 ChatGPT's identifications.
    - The prompt used to ask ChatGPT is as follows:



You

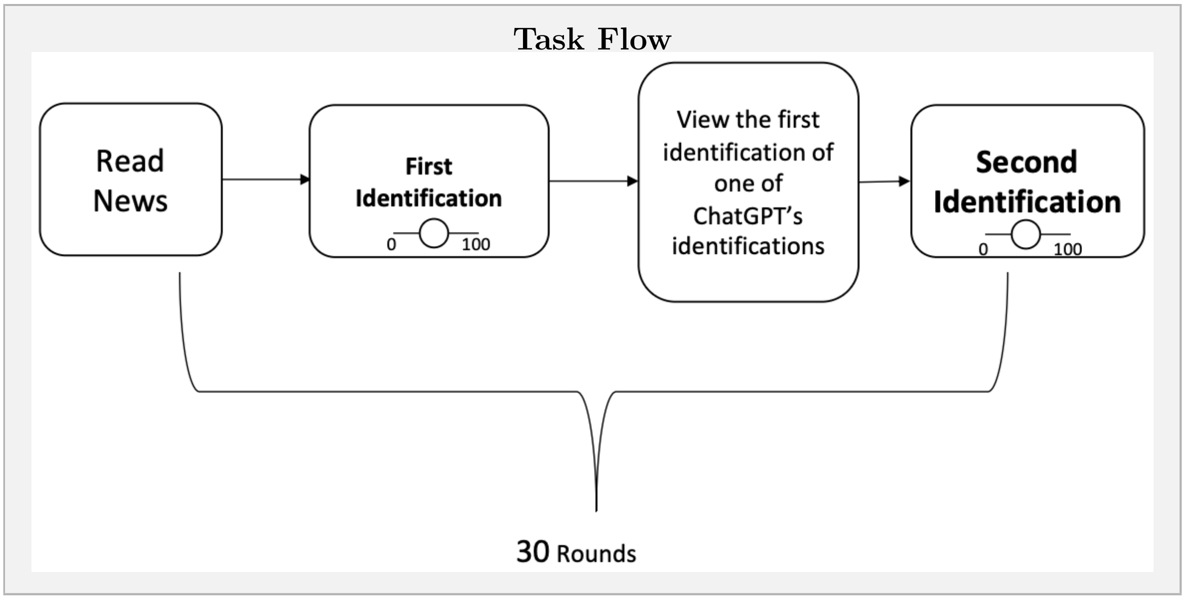
-We will now send you some Japanese news. Please identify how real it is, and report your identification as to its authenticity as an integer from 0 to 100, with 0 representing totally fake news and 100 representing totally real news.

-Do not say anything else about the result of your identification.

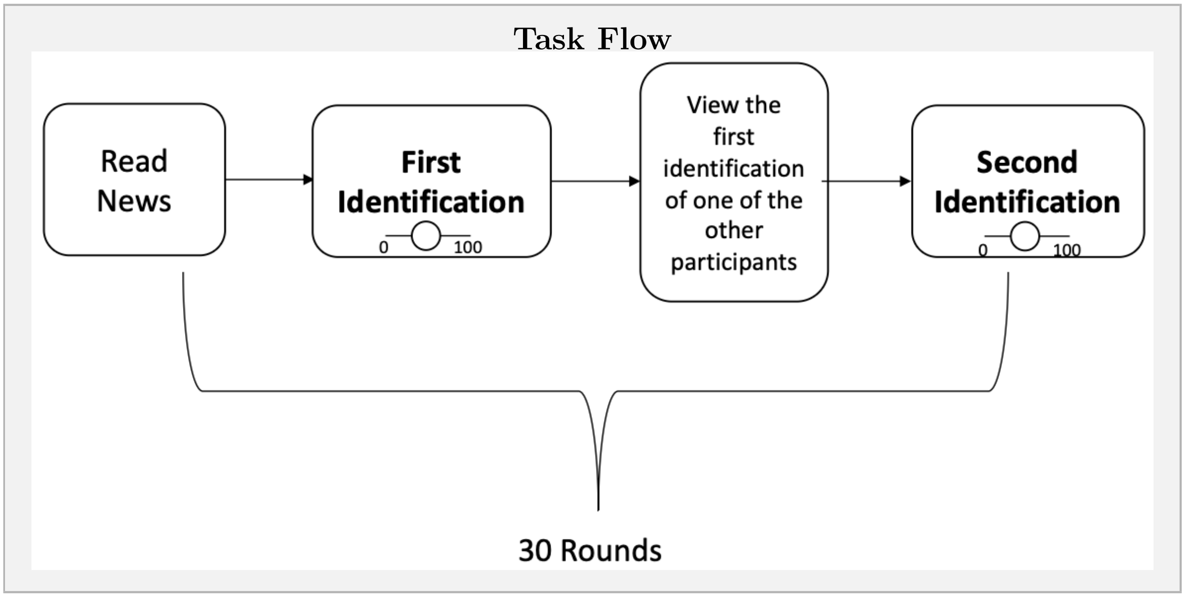
ChatGPT

Understood, please send the news when you're ready.

Treatments **AI, AIadd** only:

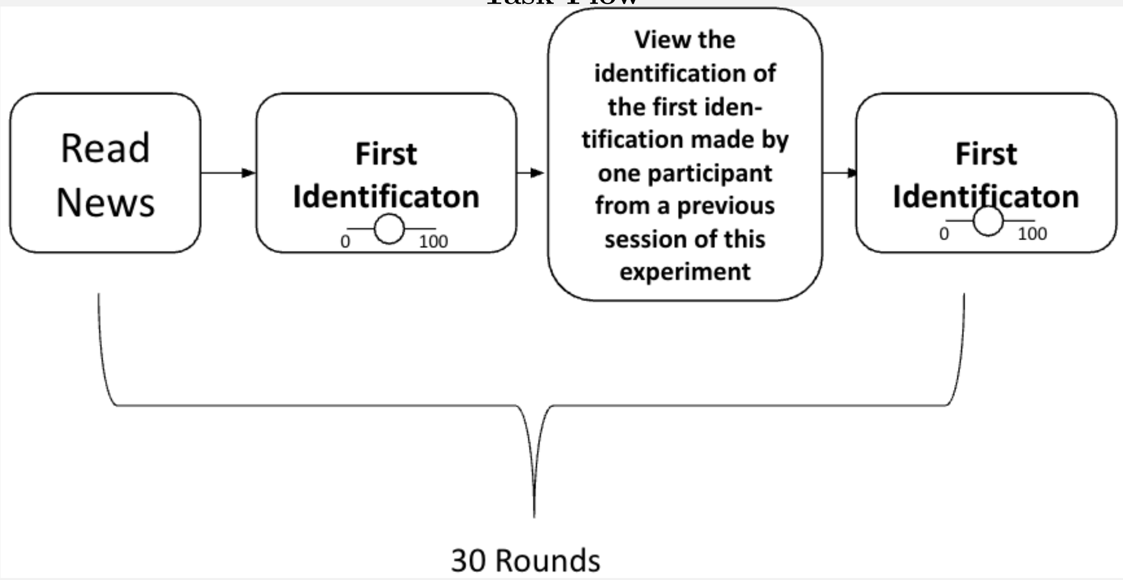


Treatment **Human** only:



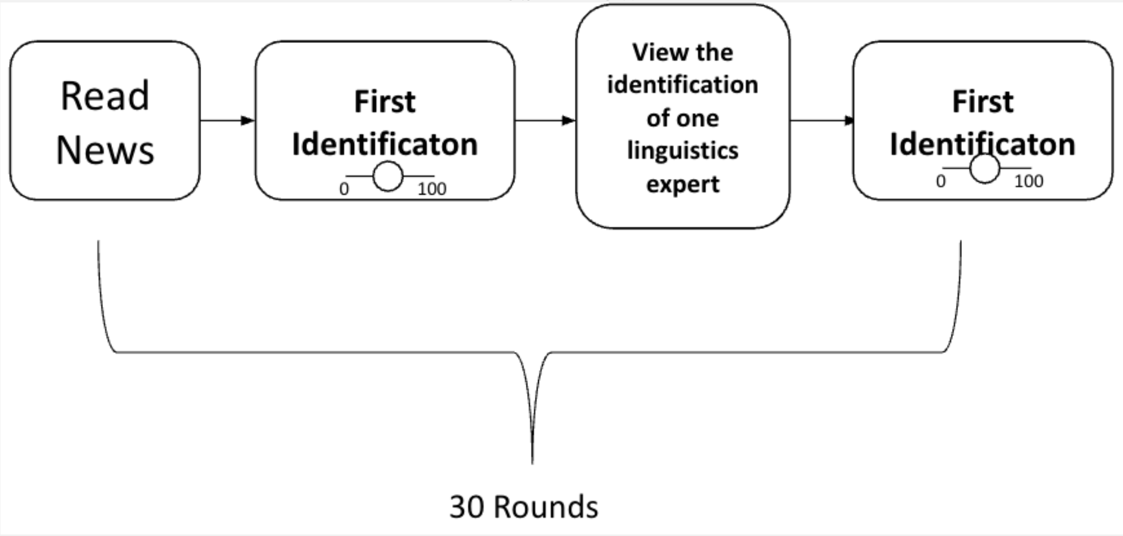
Treatment **PreHuman** only:

Task Flow


Read News
First Identificaton
0
100
View the identification of the first iden-tification made by one participant from a previous session of this experiment
First Identificaton
0
100
30 Rounds

Treatment **Expert** only:

Task Flow

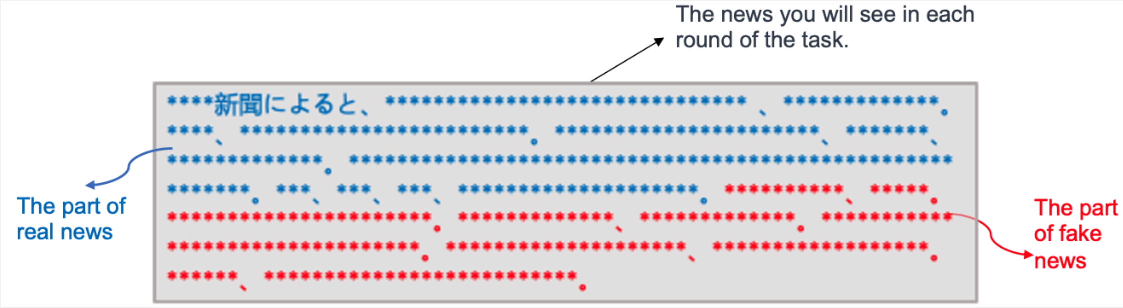


News

- The news used in the experiment is a combination of real and fake news.
  - The Real news is news written by humans and collected from Japanese wiki news.

- The Fake news is news generated by algorithms through a machine-learning model (GPT-2).

Combination of News



- Real and fake news are combined in certain proportions as shown in the figure above. The proportion of the part of real news : $real\_r$ is defined as

$$real\_r = \frac{the\ number\ of\ the\ characters\ of\ real\ news\ part}{the\ number\ of\ the\ characters\ of\ the\ whole\ news}$$

- Since the news used in the experiment includes news written totally by humans and news generated totally by algorithm, $real\_r = 100$ and $real\_r = 0$ are also possible.

*Note: The **real_r** here corresponds to $HMpro$ as defined in the main text.*

Additional Payoff

- Your additional payoff π depends on the accuracy of one randomly selected identification from all your responses throughout this experiment (a total of 30 rounds × 2 identifications= 60 responses).

- π was determined using the following equation.

$$\pi = max\{0,\ 2300 - 0.3 \times (R - real\_r)^2\}\ JPY$$

  - R : the randomly selected identification
  - $real\_r$: the proportion of the real part of the news in the selected round, after rounding off.

- In the payment of the final payoff, any fractions less than 10 yen in the final reward will be rounded up.

Note: The ***real_r*** *here corresponds to HMpro as defined in the main text.*

Quiz To check whether you understood these instructions correctly, please answer the following questions.

Please click "**Next**" button on the screen.

G Quiz Questions

There are 4 quiz questions designed to ensure that participants fully understood the experimental rules. Participants answered these questions sequentially. After each submission, they were shown whether their response was correct or incorrect, along with an explanatory comment. If a participant answered incorrectly, they were required to retry the same question until the correct answer was given; only then could they proceed to the next question.

The quiz questions, together with their answer options and explanatory comments, are presented below; the correct answers are framed.

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

- **Answer Options**
  - [✓ *Yes*]
  - *No*
- **Comments**: *"In some cases, the news is totally real or totally fake."*

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

- **Answer Options**
  - [✓ *Yes*]
  - *No*
- **Comments**: *"The true value corresponds to the proportion of real content in the news article."*

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

- **Answer Options**
  - *Yes*
  - [✓ *No*]
- **Comments**: *"The additional payment π is determined based on the accuracy of one identifications randomly selected from all of your identifications."*

Q4 *After making the first identification and receiving ChatGPT's identification, you must make a second identification different from the first one.* (Treatments **AI, AIadd** only)

Q4 *After making the first identification and receiving another participant's first identification, you must make a second identification different from the first one.* (Treatment **Human** only)

Q4 *After making the first identification and receiving the first identification of a participant from a previous session of this experiment, you must make a second identification different from the first one.* (Treatments **preHuman** only)

Q4 *After making the first identification and receiving the first identification of a linguistics expert, you must make a second identification different from the first one.* (Treatment **Expert** only)

- **Answer Options**
  - *Yes*
  - [✓ *No*]
- **Comments**: *"It does not have to be different."*

H Survey Questions

The survey questions used in our experiment are listed below. Their timing differed across experimental waves: in the main experiment (Treatments **AI** and **Human**), they were elicited before the main task, while in the additional experiment (Treatments **preHuman**, **Expert**, and **AIadd**), they were elicited after task completion.

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: *Do you have any experience with programming?* [ ]

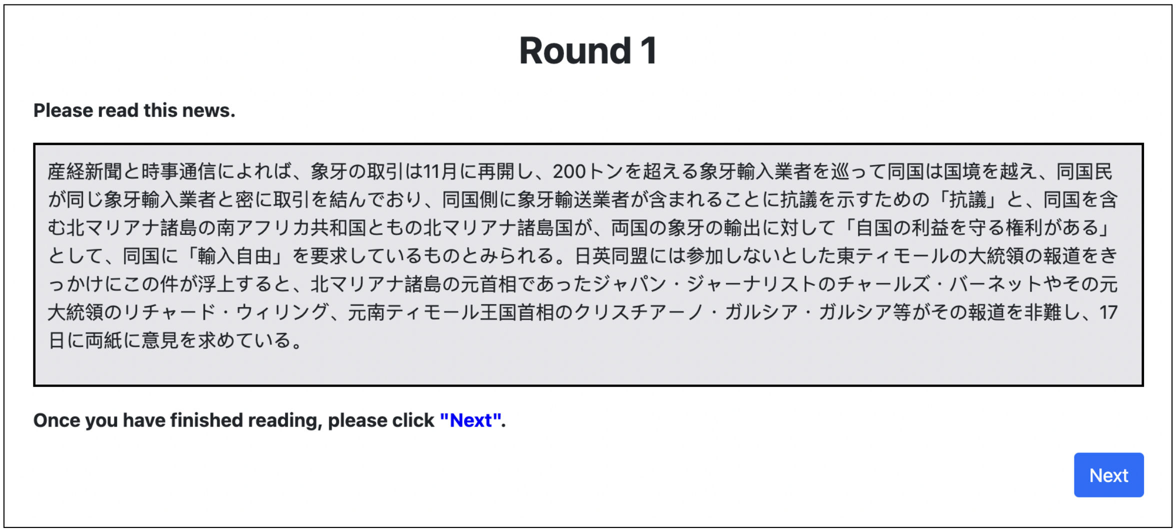
SQ5: *Have you heard about ChatGPT?* [ ]

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," which do you think can provide more accurate responses?* [Generative AI / Humans / Unsure]

I Experiment Screens

I.1 Screens of the Main Task



Round 1

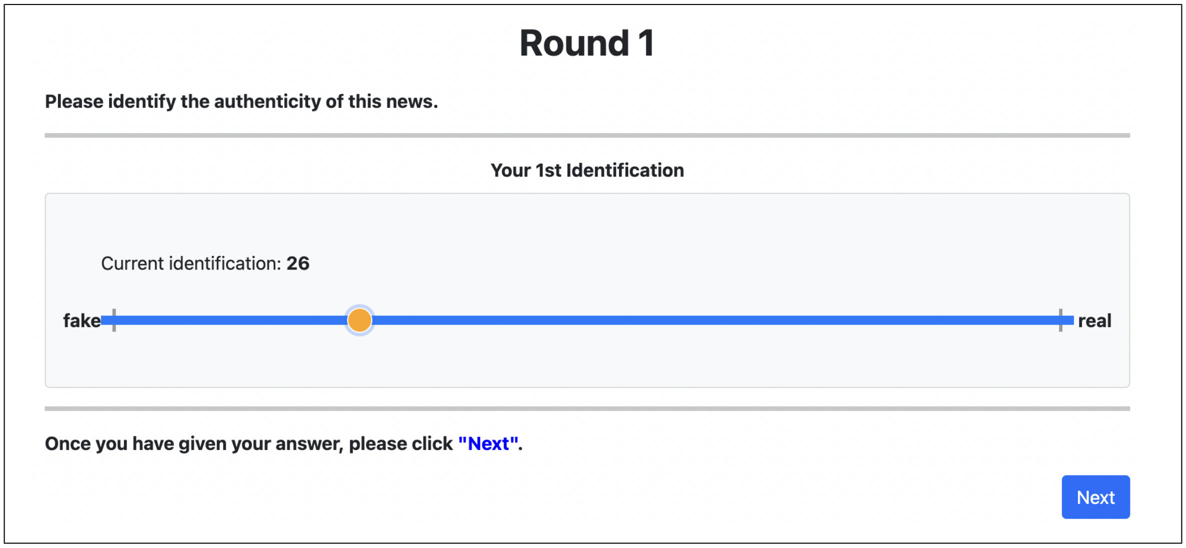
Please read this news.

産経新聞と時事通信によれば、象牙の取引は11月に再開し、200トンを超える象牙輸入業者を巡って同国は国境を越え、同国民が同じ象牙輸入業者と密に取引を結んでおり、同国側に象牙輸送業者が含まれることに抗議を示すための「抗議」と、同国を含む北マリアナ諸島の南アフリカ共和国ともの北マリアナ諸島国が、両国の象牙の輸出に対して「自国の利益を守る権利がある」として、同国に「輸入自由」を要求しているものとみられる。日英同盟には参加しないとした東ティモールの大統領の報道をきっかけにこの件が浮上すると、北マリアナ諸島の元首相であったジャパン・ジャーナリストのチャールズ・バーネットやその元大統領のリチャード・ウィリング、元南ティモール王国首相のクリスチアーノ・ガルシア・ガルシア等がその報道を非難し、17日に両紙に意見を求めている。

Once you have finished reading, please click "Next".

Figure I.1: Read News

Note: The translation of the news text: "*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.*"



Round 1

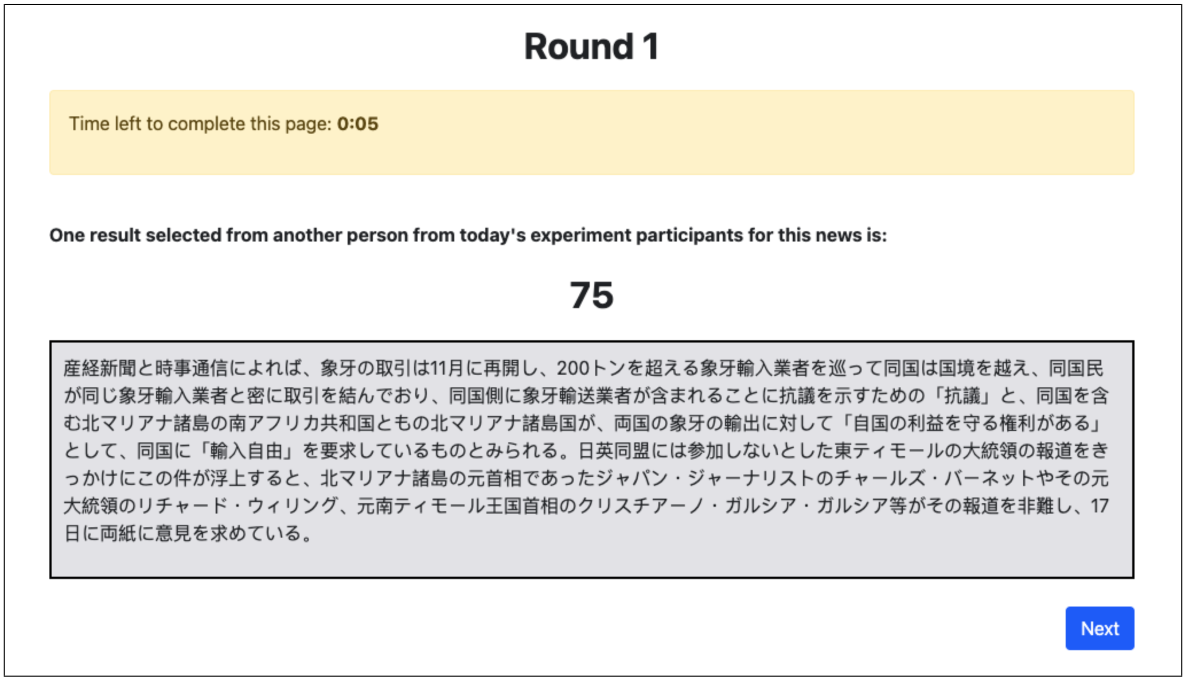
Please identify the authenticity of this news.

Your 1st Identification

Current identification: **26**

Once you have given your answer, please click "Next".

Figure I.2: First Identification



Round 1

Time left to complete this page: **0:05**

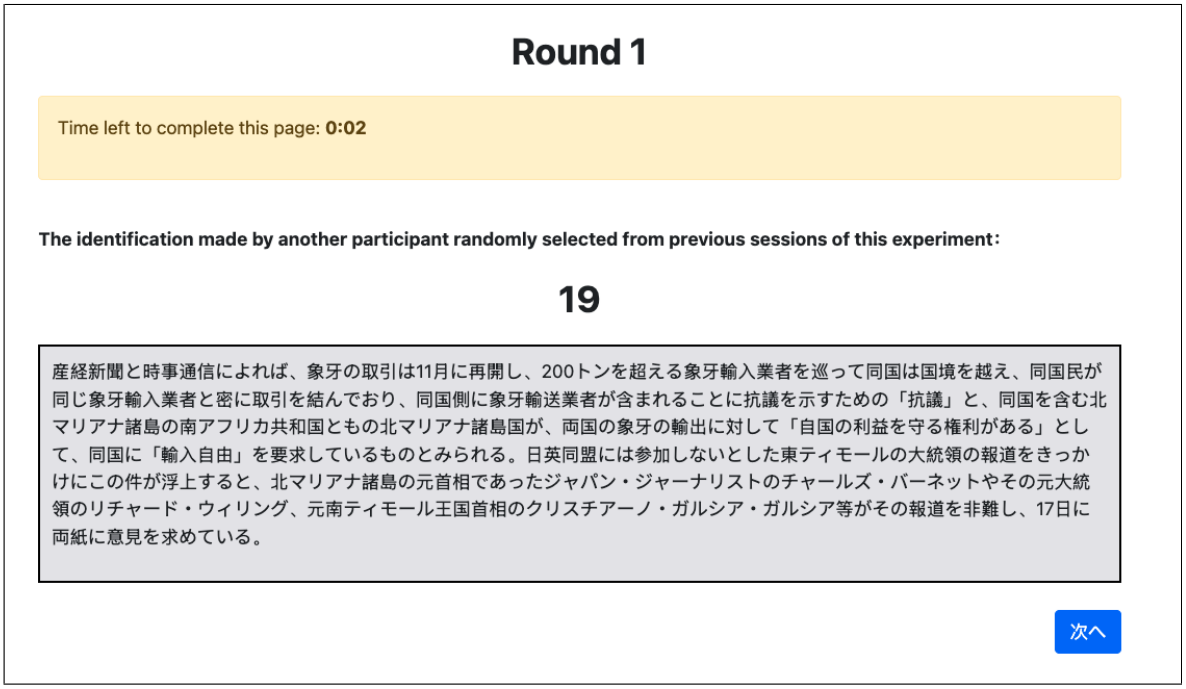
One result selected from another person from today's experiment participants for this news is:

75

産経新聞と時事通信によれば、象牙の取引は11月に再開し、200トンを超える象牙輸入業者を巡って同国は国境を越え、同国民が同じ象牙輸入業者と密に取引を結んでおり、同国側に象牙輸送業者が含まれることに抗議を示すための「抗議」と、同国を含む北マリアナ諸島の南アフリカ共和国ともの北マリアナ諸島国が、両国の象牙の輸出に対して「自国の利益を守る権利がある」として、同国に「輸入自由」を要求しているものとみられる。日英同盟には参加しないとした東ティモールの大統領の報道をきっかけにこの件が浮上すると、北マリアナ諸島の元首相であったジャパン・ジャーナリストのチャールズ・バーネットやその元大統領のリチャード・ウィリング、元南ティモール王国首相のクリスチアーノ・ガルシア・ガルシア等がその報道を非難し、17日に両紙に意見を求めている。

Figure I.3: Advice Display (Treatment **Human** only)

Note: The translation of the news text: "*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.*"



Round 1

Time left to complete this page: **0:02**

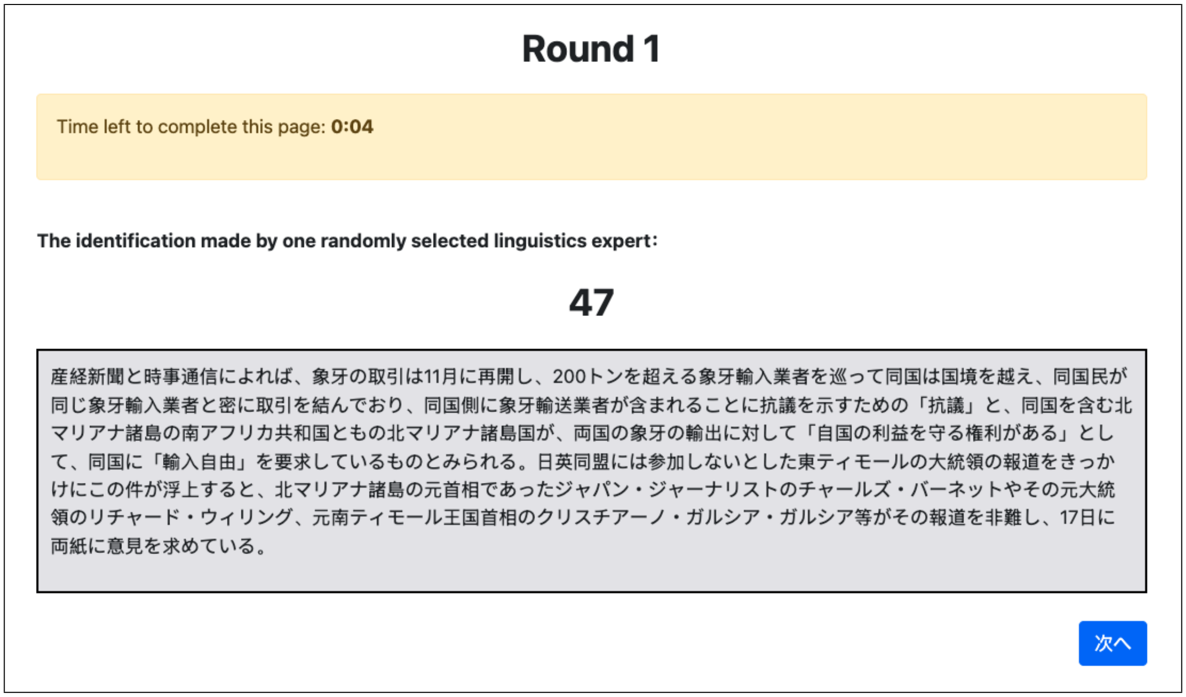
The identification made by another participant randomly selected from previous sessions of this experiment:

19

産経新聞と時事通信によれば、象牙の取引は11月に再開し、200トンを超える象牙輸入業者を巡って同国は国境を越え、同国民が同じ象牙輸入業者と密に取引を結んでおり、同国側に象牙輸送業者が含まれることに抗議を示すための「抗議」と、同国を含む北マリアナ諸島の南アフリカ共和国ともの北マリアナ諸島国が、両国の象牙の輸出に対して「自国の利益を守る権利がある」として、同国に「輸入自由」を要求しているものとみられる。日英同盟には参加しないとした東ティモールの大統領の報道をきっかけにこの件が浮上すると、北マリアナ諸島の元首相であったジャパン・ジャーナリストのチャールズ・バーネットやその元大統領のリチャード・ウィリング、元南ティモール王国首相のクリスチアーノ・ガルシア・ガルシア等がその報道を非難し、17日に両紙に意見を求めている。

Figure I.4: Advice Display (Treatment **preHuman** only)

Note: The translation of the news text: "*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.*"



Round 1

Time left to complete this page: **0:04**

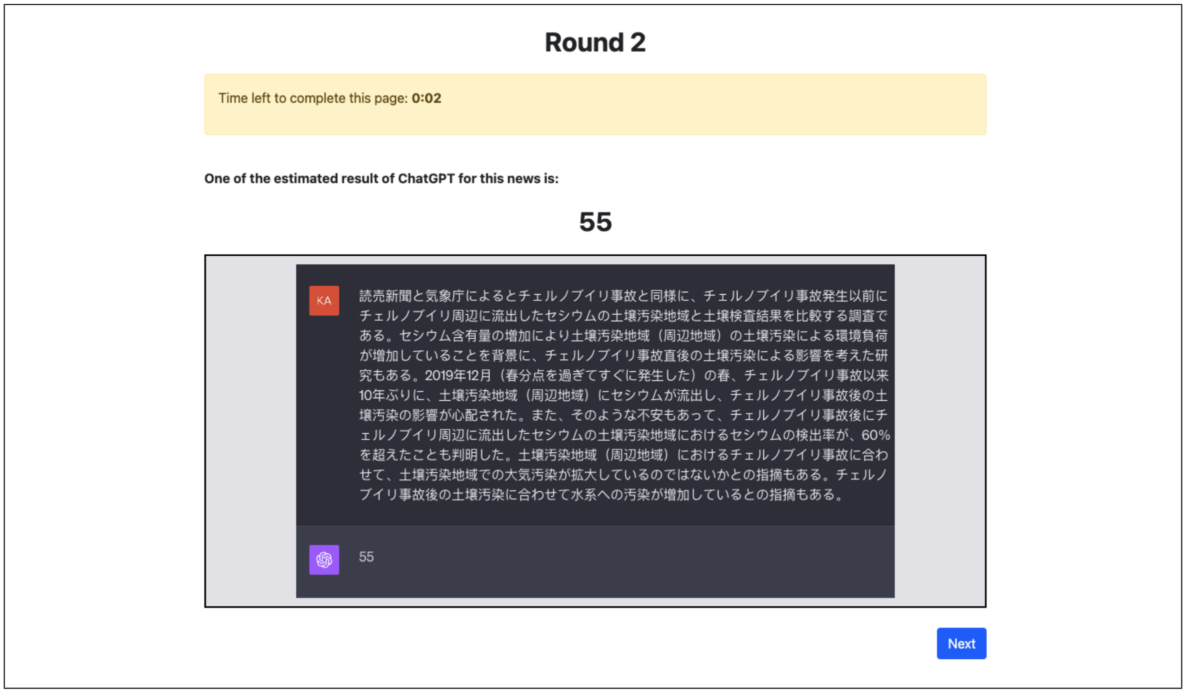
The identification made by one randomly selected linguistics expert:

47

産経新聞と時事通信によれば、象牙の取引は11月に再開し、200トンを超える象牙輸入業者を巡って同国は国境を越え、同国民が同じ象牙輸入業者と密に取引を結んでおり、同国側に象牙輸送業者が含まれることに抗議を示すための「抗議」と、同国を含む北マリアナ諸島の南アフリカ共和国ともの北マリアナ諸島国が、両国の象牙の輸出に対して「自国の利益を守る権利がある」として、同国に「輸入自由」を要求しているものとみられる。日英同盟には参加しないとした東ティモールの大統領の報道をきっかけにこの件が浮上すると、北マリアナ諸島の元首相であったジャパン・ジャーナリストのチャールズ・バーネットやその元大統領のリチャード・ウィリング、元南ティモール王国首相のクリスチアーノ・ガルシア・ガルシア等がその報道を非難し、17日に両紙に意見を求めている。

Figure I.5: Advice Display (Treatment **Expert** only)

Note: The translation of the news text: "*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.*"



Round 2

Time left to complete this page: **0:02**

One of the estimated result of ChatGPT for this news is:

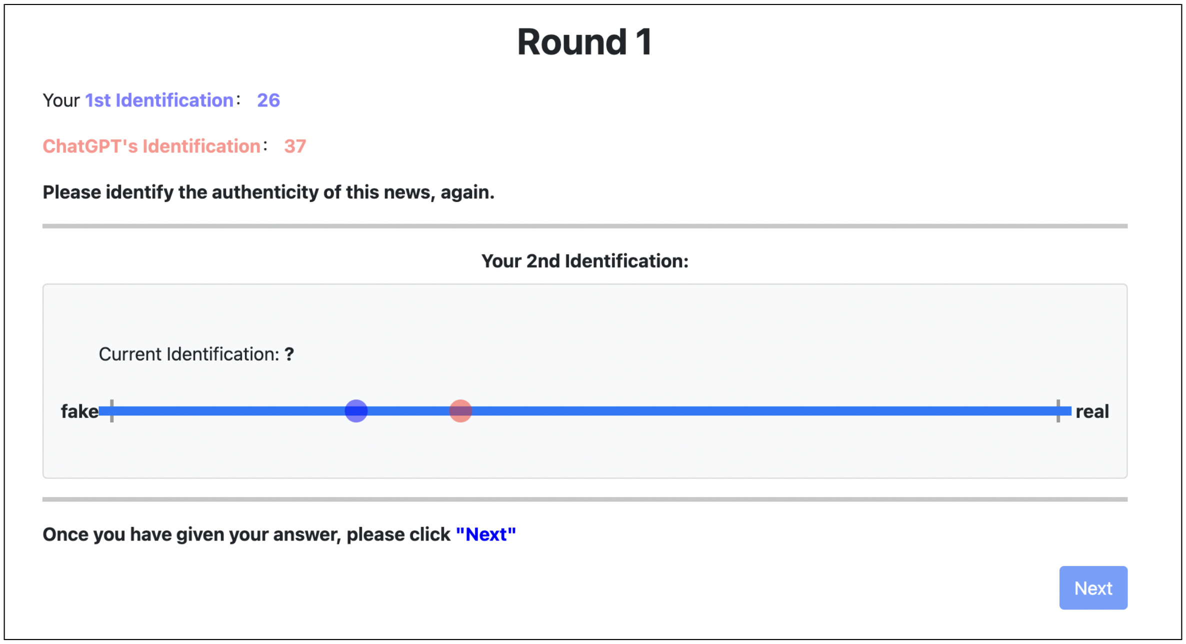
55

読売新聞と気象庁によるとチェルノブイリ事故と同様に、チェルノブイリ事故発生以前にチェルノブイリ周辺に流出したセシウムの土壌汚染地域と土壌検査結果を比較する調査である。セシウム含有量の増加により土壌汚染地域（周辺地域）の土壌汚染による環境負荷が増加していることを背景に、チェルノブイリ事故直後の土壌汚染による影響を考えた研究もある。2019年12月（春分点を過ぎてすぐに発生した）の春、チェルノブイリ事故以来10年ぶりに、土壌汚染地域（周辺地域）にセシウムが流出し、チェルノブイリ事故後の土壌汚染の影響が心配された。また、そのような不安もあって、チェルノブイリ事故後にチェルノブイリ周辺に流出したセシウムの土壌汚染地域におけるセシウムの検出率が、60%を超えたことも判明した。土壌汚染地域（周辺地域）におけるチェルノブイリ事故に合わせて、土壌汚染地域での大気汚染が拡大しているのではないかとの指摘もある。チェルノブイリ事故後の土壌汚染に合わせて水系への汚染が増加しているとの指摘もある。

55

Figure I.6: Advice Display (Treatments **AI, AIadd** only)

Note: The translation of the news text: "*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.*"



Round 1

Your 1st Identification：26

ChatGPT's Identification：37

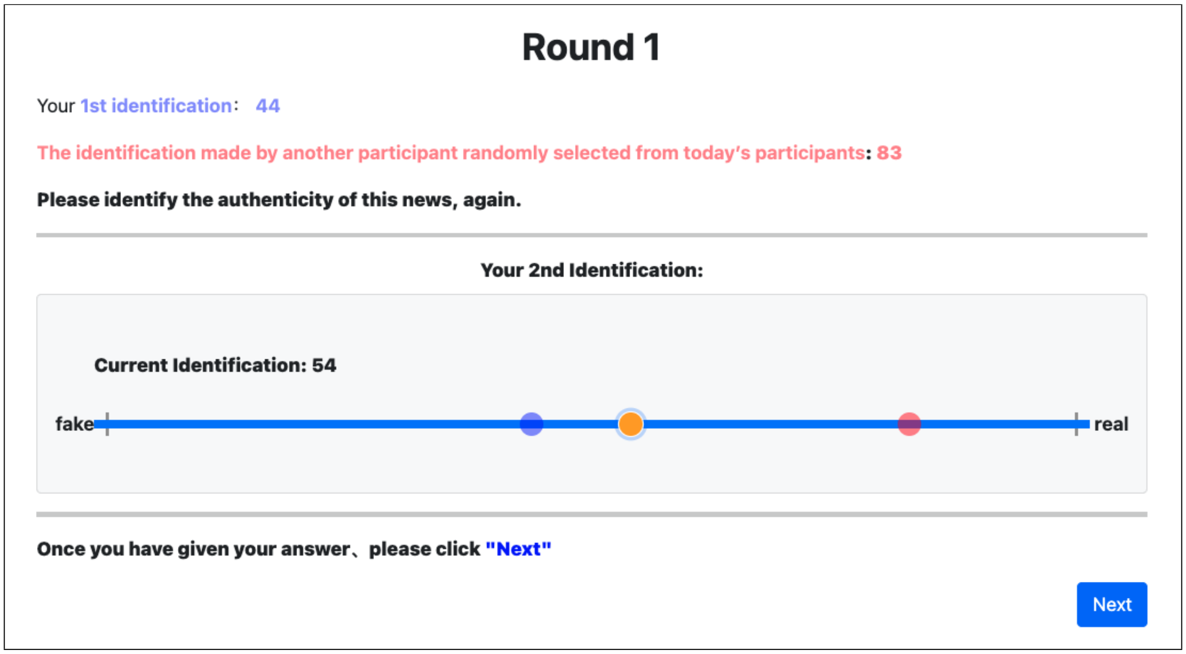
Please identify the authenticity of this news, again.

Your 2nd Identification:

Current Identification: ?

Once you have given your answer, please click "Next"

Figure I.7: Second Identification (Treatments **AI, AIadd** only)



Round 1

Your 1st identification：44

The identification made by another participant randomly selected from today's participants: 83

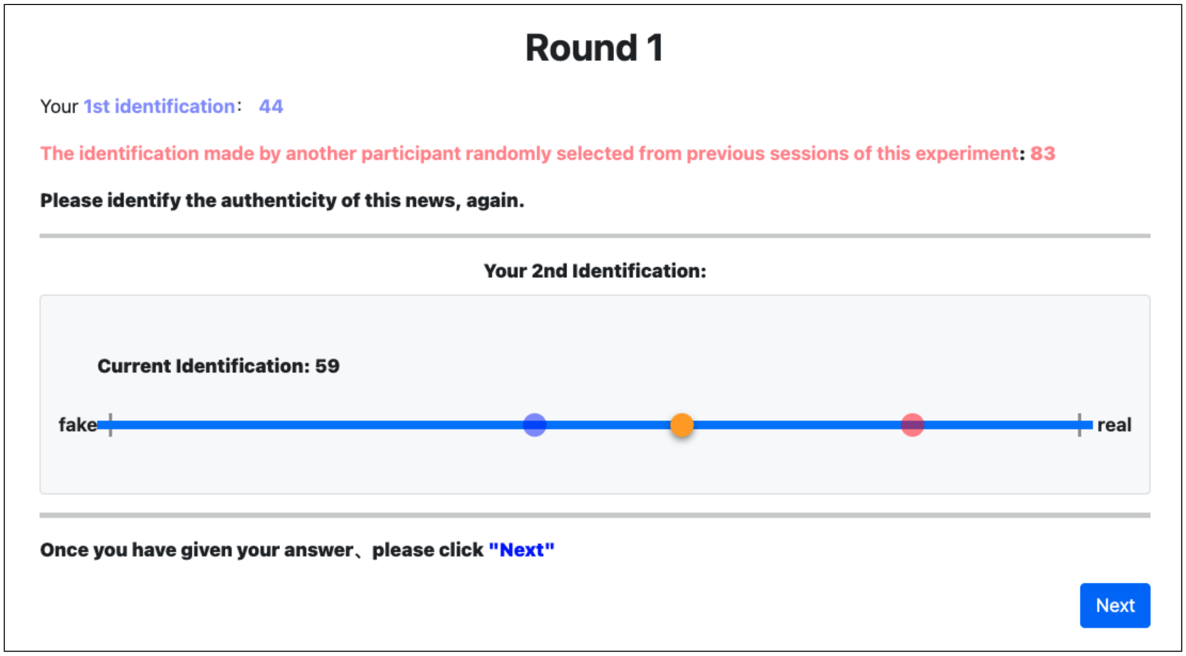
Please identify the authenticity of this news, again.

Your 2nd Identification:

Current Identification: 54

Once you have given your answer、please click "Next"

Figure I.8: Second Identification (Treatment **Human** only)



Round 1

Your 1st identification： 44

The identification made by another participant randomly selected from previous sessions of this experiment: 83

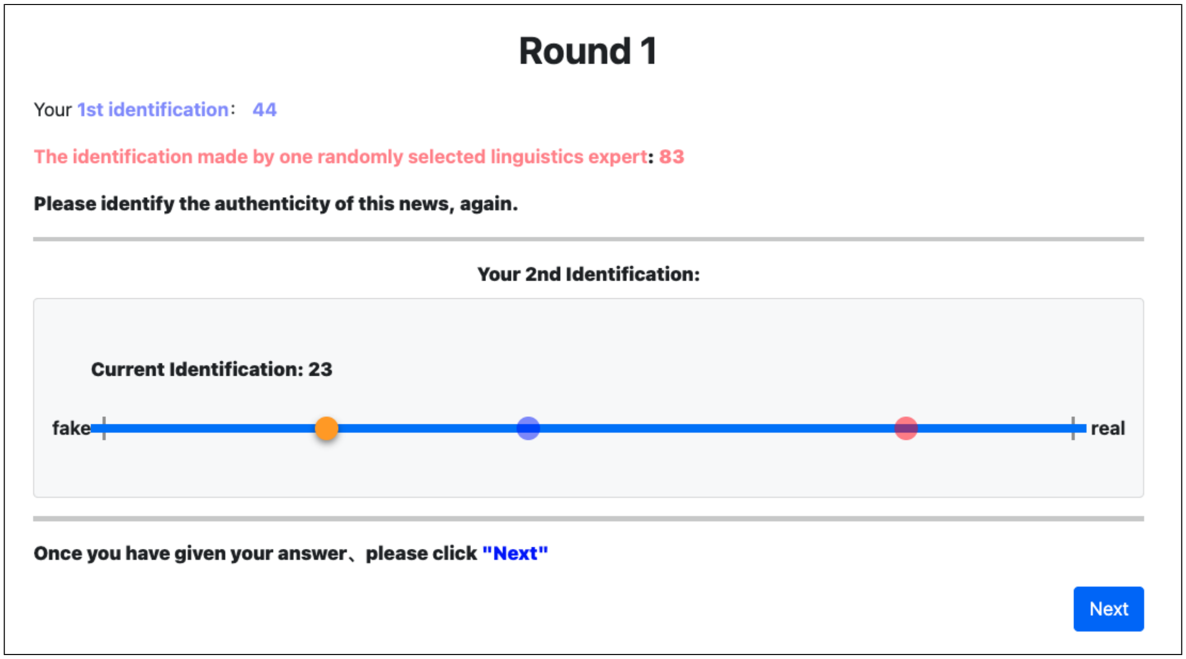
Please identify the authenticity of this news, again.

Your 2nd Identification:

Current Identification: 59

Once you have given your answer、please click "Next"

Figure I.9: Second Identification (Treatment **preHuman** only)



Round 1

Your 1st identification： 44

The identification made by one randomly selected linguistics expert: 83

Please identify the authenticity of this news, again.

Your 2nd Identification:

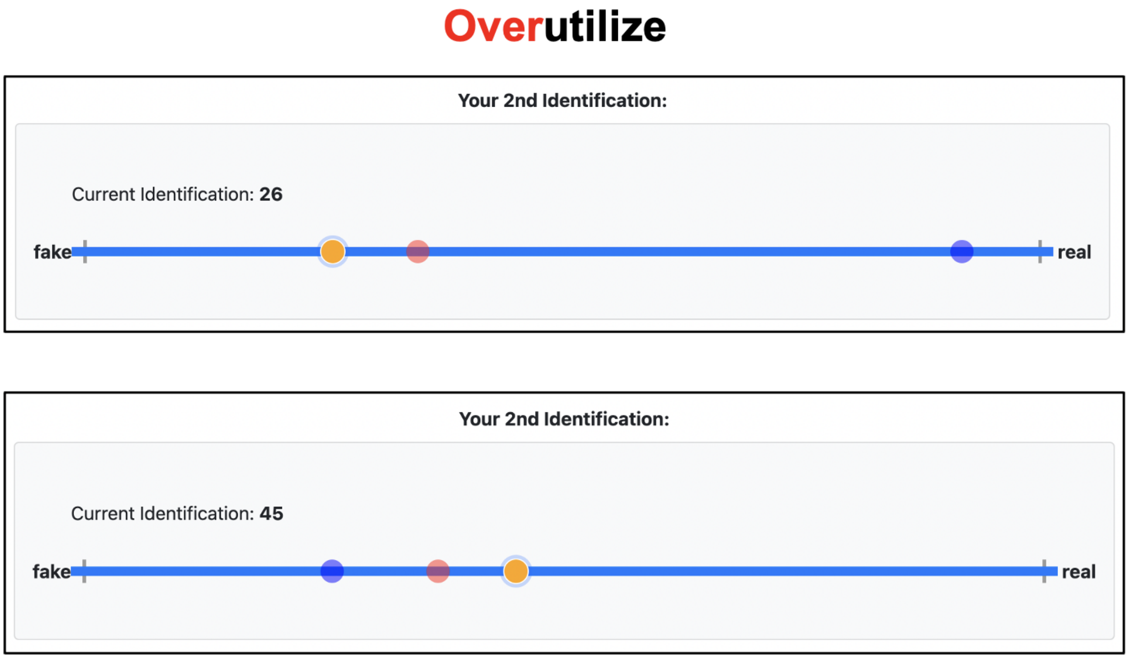
Current Identification: 23

Once you have given your answer、please click "Next"

Figure I.10: Second Identification (Treatment **Expert** only)

I.2 Examples of Utilization Statuses

The following plots show examples of the three utilization statuses described in Section C.2. 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.



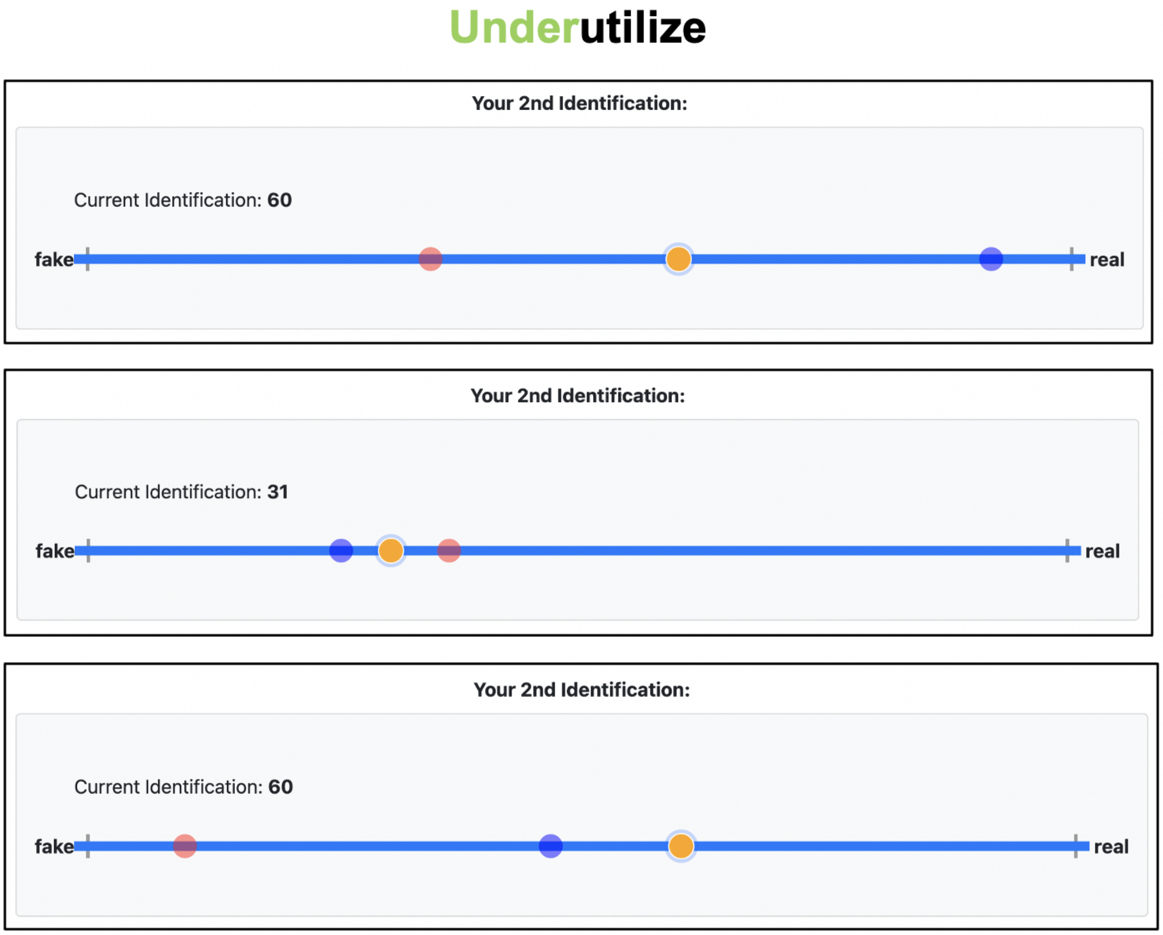
Your 2nd Identification:

Current Identification: **26**

Your 2nd Identification:

Current Identification: **45**

Figure I.11: Two Examples of Over-utilize



Your 2nd Identification:

Current Identification: **60**

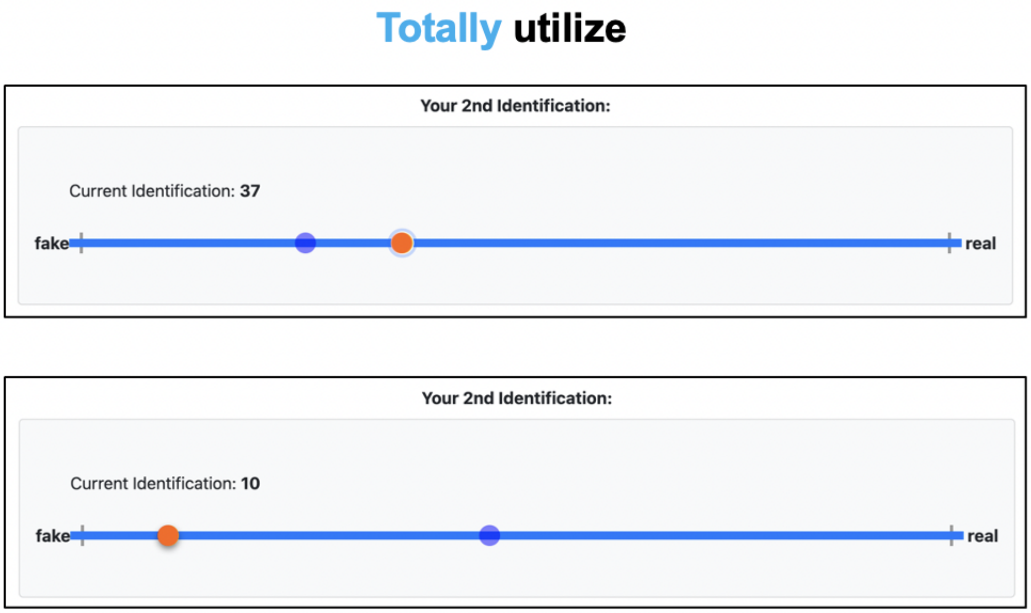
Your 2nd Identification:

Current Identification: **31**

Your 2nd Identification:

Current Identification: **60**

Figure I.12: Three Examples of Under-utilize



Totally utilize

Your 2nd Identification:

Current Identification: **37**

Your 2nd Identification:

Current Identification: **10**

Figure I.13: Two Examples of Totally-utilize

References

J. I. Baines, R. S. Dalal, L. P. Ponce, and H.-C. Tsai. Advice from artificial intelligence: a review and practical implications. *Frontiers in Psychology*, 15:1390182, 2024.

P. Ecken and R. Pibernik. Hit or miss: What leads experts to take advice for long-term judgments? *Management Science*, 62(7):2002–2021, 2016.

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

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

M. Hütter and F. Ache. Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, 11(4):401–415, 2016.

N. Mesbah, C. Tauchert, and P. Buxmann. Whose advice counts more–man or machine? an experimental investigation of AI-based advice utilization. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, page 496, 2021. doi: 10.24251/hicss.2021.496.

M. Moussaïd, J. E. Kämmer, P. P. Analytis, and H. Neth. Social influence and the collective dynamics of opinion formation. *PloS one*, 8(11):e78433, 2013.

T. Schultze, A.-F. Rakotoarisoa, and S.-H. Stefan. Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision making*, 10(2): 144–171, 2015.

K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou. 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*, pages 763–777, 2022.