Discussion Paper No. 1296

ISSN (Print) 0473-453X
ISSN (Online) 2435-0982

PAYING AI TO DETECT AI

Yuhao Fu
Nobuyuki Hanaki

November 2025

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

Paying AI to Detect AI*

Yuhao Fu† Nobuyuki Hanaki‡

November 10, 2025

Abstract

We embed a ChatGPT-based AI detector in a laboratory experiment to test whether participants are willing to pay more to collaborate with AI than with human peers to accurately detect the proportion of AI-generated parts in deepfake news articles. Task difficulty varies with the model used to generate the articles (GPT-2 vs. GPT-4o). We find that participants' willingness to pay (WTP) for the AI detector exceeds that to collaborate with human peers, even though the AI detector does not provide better assistance and, in fact, humans do better than AI in for GPT-4o generated news. WTP for AI or peer collaboration does not rise with task difficulty. These patterns point to over-reliance on AI and raise concerns about the spread of deepfakes. The study improves understanding of human–AI interaction and informs safeguards for deepfake detection in the era of generative AI (GAI).

Keywords: ChatGPT; AI detector; human–AI collaboration; willingness to pay; AI reliance; deepfake detection

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), the University of Osaka, and (b) Grants-in-aid for Scientific Research (No.23H00055 and No.25H00388) from the Japan Society for the Promotion of Science. The design of the experiment reported in this paper was approved by the Research Ethics Committee at the ISER (#20240904) and the Research Ethics Committee at the Research Institute for Socionetwork Strategies (RISS), Kansai University (#2024030). The experiment is preregistered at aspredicted.org (#208604). Support from Yuta Shimodaira, Satsuki Yamada, and staff of RISS for conducting the experiment is gratefully acknowledged.

†Graduate School of Economics, the University of Osaka. E-mail: kazewindser@gmail.com

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

1 Introduction

With the popularity of AI tools since the launch of ChatGPT in November 2022, the spread of AI-generated fake contents (deepfakes)– has also raised concerns on its harm to the Human society (Lundberg and Mozelius, 2024; Sophia, 2025). Recognizing this, "AI detectors" (e.g., Turnitin) have been developed (Kar et al., 2024) and are used for academic integrity checks and for fact-checking on 𝕏 (e.g., TweetDetective). Beyond education, platforms label AI-generated media using an open content-provenance standard for content credentials (e.g., Meta, TikTok); firms deploy deepfake detectors against voice and multimedia fraud; and consumer tools let ordinary users test images, audio, video, and text.[1]

However, AI detectors are imperfect and can misclassify human writing as AI-generated (false positives) (Weber-Wulff et al., 2023; Knilans, 2024; Zhang et al., 2024b). Even OpenAI advises that AI-writing detectors are not reliable for high-stakes judgments.[2] Well-known cases include detectors labeling the *US Constitution* and passages from the *Bible* as AI-generated,[3] and false positives have led to real harms in education, including publicized misconduct cases and delayed graduations (Bergin, 2025b,a; Hallikaar, 2025; The Advertiser, 2025; The Courier-Mail, 2024).

With this context in mind, this study seeks to quantify people's trust in these imperfect AI tools. We introduced a deepfake detection task in the laboratory, where participants were asked to identify the proportion of AI-generated parts in deepfake news articles, which were compositions of human-written and AI-generated text. Using a between-subjects design, we then offered paid external assistance: participants may either (i) purchase access to a ChatGPT-based AI detector or (ii) pay to cooperate

[1] Examples include Turnitin's one-year AI-detector report; Meta/TikTok updates on the content-credentials standard (also known as C2PA); Pindrop and Reality Defender for enterprise; and consumer tools such as AI or Not, Google's SynthID, and GPTZero.

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

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

with other participants in the experiment before revising their initial identifications. We utilize the Becker–DeGroot–Marschak (BDM) mechanism (Becker et al., 1964) to elicit participants' WTP for the external assistance and our first research question is therefore

RQ1: *Are participants willing to pay more to use ChatGPT than to cooperate with other participants to detect deepfake news?*

As generative AI (GAI) technologies advance, a central goal is to produce outputs that appear ever more natural and human-like. This progress makes deepfakes increasingly difficult for the public to recognise and imposes additional burdens on detection systems, thereby magnifying their potential harm (Masood et al., 2023; Lee, 2024). To examine how this technological progress affects both deepfake detection and reliance on AI detectors, we introduce the second dimension in our experiment that systematically varies task difficulty. Concretely, we construct two sets of synthetic news articles using different large language models (LLMs)—GPT-2 and GPT-4o—so that detecting AI-generated content in the GPT-4o set is theoretically more challenging (Islam and Moushi, 2024). This leads to our second research question:

RQ2: *Are participants willing to pay more for external assistance (using ChatGPT or cooperating with Human peers) when facing more difficult tasks (detecting GPT-4o generated news) compared to less difficult tasks (GPT-2 generated news)?*

Many GAI chatbots—such as ChatGPT—display disclaimers that their outputs may contain errors, discouraging blind trust. Tools built on the same technology—AI detectors—rarely provide comparable notices about their own fallibility or their non-negligible false-positive rates; related policies to limit misuse are also scarce. Without systematic validation or feedback, users' choices to employ these detectors—and to believe their results—remain largely subjective. We test whether exposing users to evidence of a tool's unreliability can reduce over-trust in detectors, as seen for chatbots.

In a laboratory experiment, participants complete two successive deepfake-detection parts. Before each part, they submit their WTP for the "external assistance" available in the upcoming rounds. After the first part, a mid-experiment feedback stage provides performance feedback. This design allows us to observe whether experience and feedback about a given form of assistance change subsequent demand for—and reliance on—that assistance. We therefore pose our third research question:

RQ3: *Do the experience of external assistance (using ChatGPT or cooperating with Human peers) and the feedback regarding its effectiveness in detecting deepfake news affect their willingness to pay (WTP) to use them?*

Despite rapid growth in the deepfake literature, few economic-experiment studies combine a deepfake-detection task with an AI detector, as we do here. "AI reliance" (Passi and Vorvoreanu, 2022; Klingbeil et al., 2024; Küper and Krämer, 2025) becomes problematic in this context when two conditions coincide: (i) participants place more trust in an AI detector than in human collaboration, and (ii) the detector does not deliver superior assistance compared to humans. In other words, people's trust in AI tools is excessive—and potentially harmful—when the tool commands more confidence than its performance warrants. To assess the objective value of AI assistance (rather than the value users subjectively place on it), we pose our fourth research question:

RQ4: *Does ChatGPT help people more effectively than Human peers in the tasks of deepfake detection?*

Our findings show that participants detect GPT-4o deepfake news less accurately than GPT-2 deepfake news, confirming that the task is harder when the generating model is stronger. Participants' WTP is higher for the AI detector than for collaboration with human peers, and this premium remains essentially unchanged across difficulty levels. After the first part and performance feedback, demand rises further: participants are more likely to increase their WTP for access to the AI detector. Finally, the

AI detector does not improve detection accuracy relative to peer collaboration, even though participants believe it does and pay a premium for it.

These findings call for policy frameworks that help people and institutions evaluate AI detectors prudently when facing the risks of deepfakes. Regulators should require transparent disclosure of performance metrics—not only for generative-AI systems but also for detectors—to reduce over-reliance on AI tools and mitigate potential harm. A complementary step is to establish an independent body to test detectors against common standards, and to require platforms to use a human reviewer for any content that a detector flags with high confidence as AI-generated, rather than leaving the decision to AI alone. Finally, policies that support the development of detectors with higher accuracy, lower false-positive rates, and fairer pricing can help keep reliable detection in step with the growing sophistication of GAI.

The remainder of this paper is organized as follows. Section 2 introduces the background of deepfake detection and reviews previous studies on valuing the Human-AI Collaboration. Section 3 presents the experimental design and hypotheses. Section 4 summarizes the analysis results. Section 5 provides a discussion, and Section 6 concludes the paper.

2 Literature Review

We review two strands of research. First, we summarize the state of deepfake detection—its motivations, main approaches, and key challenges. Second, we survey work on valuing human–AI collaboration, with attention to motivations, measurement methods, and comparisons to human–human collaboration.

2.1 Background on Deepfake Detection

2.1.1 Why deepfakes matter

Deepfakes are synthetic media—most often audio and video—created with deep learning methods (Chadha et al., 2021). These outputs imitate real content closely and can cause social harm (Katarya and Lal, 2020; Sareen, 2022). With recent advances in GAI, the idea of deepfakes now extends beyond images and audio to include text (Chong et al., 2023). In particular, AI-generated deepfake news—"fake" text that looks credible—has drawn serious concern (Lee and Shin, 2022; Guo, 2024). As LLMs improve, deepfakes have become harder to distinguish from real content, and their potential harms across domains have drawn increased attention.

Politically, deepfakes are seen as a major threat to political processes. Islam et al. (2024) argue that they can sway voter decisions and affect elections and democracy worldwide. Amin et al. (2025) show that exposure to deepfake images increases cognitive load and confirmation bias, shaping perceived political ideology and fostering polarization. Li (2025) report that deepfake news spreads at unprecedented speed and scale, appears highly authentic, and contributes to crises of social trust, political polarization, and economic and legal risks. Gupta et al. (2025) find that deepfakes shape perceptions and narratives, especially during elections and periods of political turmoil. These risks motivate both governance and user-facing safeguards beyond detection alone.

Academia and education face integrity risks from deepfakes. In research, deep learning can generate realistic but nonexistent data and images, which have already led to journal retractions and signal a systemic threat to research integrity (Chen et al., 2024). Chauhan and Currie (2024) argue that GAI raises authenticity risks in scholarly writing, including hallucinated content and citations, fabricated or "synthetic" data and images, and reference manipulation, thereby undermining reproducibility and verification. In higher education, Bittle and El-Gayar (2025) note that while GAI offers learning benefits, its deepfakes can increase opportunities for academic misconduct,

making governance and assessment redesign pressing issues. Deepfakes can also harm students through "cyber-bullying": schools face challenges in detection, reporting, and platform coordination, suggesting that current anti-bullying procedures are not ready for AI-driven harassment (Alexander, 2025).

In everyday life, deepfakes erode privacy and reputation. Non-consensual synthetic intimate imagery (NSII, i.e., deepfake pornography) has non-trivial prevalence and causes lasting psychosocial harm (Umbach et al., 2024). Cloned voices fuel phone scams because users struggle to detect them, with low accuracy for both known and unknown speakers (Barrington et al., 2025). In markets, deepfake advertising distorts consumer judgment when disclosure is weak; product claims, the presence of disclosure, and the form of disclosure all shape evaluations and trust (Whittaker et al., 2025). Overall, exposure to deepfakes can change beliefs, memories, and sharing behavior in daily media use (Ching et al., 2025).

Taken together, the wide-ranging risks of deepfakes have sharpened public concern about how to detect them. On the policy side, scholars call for governments to recognize the harms and spread of deepfakes and to adopt targeted legal responses (Yamaoka-Enkerlin, 2019; Ramluckan, 2024). On the technical side, new methods and detection tools are still needed (Mirsky and Lee, 2021). On the human side, because deepfakes are increasingly hard to tell from real content, users express a clear demand to verify media in everyday social contexts (Ahmed and Chua, 2023).

In this study, we focus on a common form of deepfake—deepfake news—and, in a laboratory setting, examine a detection task in which participants identify the proportion of AI-generated text in each article. While this design cannot fully recreate real-world contexts, it provides a controlled way to measure perceptions and behavior toward deepfakes. To our knowledge, this is among the first economic experiments to elicit WTP for deepfake detection. Our results clarify how people value and rely on detection tools and inform efforts to mitigate the harms of deepfakes.

2.1.2 Deepfake Detection and its' Frictions

Many studies show that humans can not detect deepfakes effectively. Diel et al. (2024) systematically reviewed and meta-analyzed 56 papers on human performance and found an overall detection accuracy of 55.54% (52.00% for deepfake text), providing the first comprehensive review of human deepfake detection. Groh et al. (2024) ran randomized experiments on real versus deepfake political speeches and found that the proportion of deepfakes did not significantly affect judgments; they also report that, compared to voice and video, deepfake text is harder for humans to detect.

At the same time, Human-Human Collaboration helps humans' deepfake detection. Uchendu et al. (2023) shows that group discussion improves deepfake-text detection relative to individuals, with gains for both non-experts and experts. Groh et al. (2022) demonstrates that aggregating multiple human judgments on deepfake videos yields accuracy comparable to that of state-of-the-art detectors and surpasses that of individual raters. These collaborative gains are consistent with evidence summarized in Diel et al. (2024)'s systematic reviews.

Beyond human detection, AI-based detection has become a primary approach to spotting deepfakes (Garde et al., 2022). As Zellers et al. (2019) put it, "the best way to detect neural fake news is to use a model that is also a generator." Evidence on AI detectors is mixed: some evaluations report very high performance—often above 90% and in some cases near 100%—on specific benchmarks (Koka et al., 2024; Sallami et al., 2024; Liu et al., 2024), yet other studies find off-the-shelf tools far from perfect. Weber-Wulff et al. (2023) assessed 12 public detectors and two commercial systems (Turnitin and PlagiarismCheck) and concluded that available tools are neither accurate nor reliable, with detection of AI text often only slightly above chance and vulnerable to paraphrasing—limiting their evidentiary value. Reviewing 17 articles, Chaka (2024) likewise report inconsistent results across detectors and datasets, indicating limited reliability. As LLMs improve, deepfakes become harder to detect, while detectors also

advance—an ongoing "generation–detection" arms race (Laurier et al., 2024).

Even though AI detectors are not perfect, human–AI collaboration is considered promising for deepfake detection (Saharan et al., 2025). Recent evidence shows clear gains: in experiments by Groh et al. (2022), participants who could see a model's prediction were more accurate than either humans or the model alone; Diel et al. (2024) likewise report that AI assistance can improve human detection accuracy in their review; and Somoray et al. (2025) show that humans and AI models attend to different cues when detecting deepfakes, indicating complementary strengths that collaboration can exploit.

In this study, we introduce two collaboration mechanisms—human–human and human–AI. Participants can pay to collaborate with a peer or with an AI detector when detecting deepfake news. Because many public AI detectors are now paywalled and prior work rarely measures WTP to use an AI detector for deepfake detection, this study fills that gap.

Additionally, in many experimental studies on fake-news detection in economics, participants are asked for a binary response (real vs. fake) (Serra-Garcia and Gneezy, 2021; Arin et al., 2023; Thaler, 2024); by contrast, we ask participants to identify the article's AI-generated proportion, aligning with recent work on partial detection and localization rather than whole-document labels (Zhang et al., 2024a; Zeng et al., 2024; Zhang et al., 2024d). We prefer a proportion for four reasons: (i) our stimuli include totally AI-generated and totally human-written items, so a proportion subsumes binary judgments and provides finer measurement; (ii) as models improve, simple yes/no human detection is unreliable, whereas a proportion captures subjective uncertainty; (iii) real-world use often mixes AI output with human edits (AI–human compositions), so proportion better matches how content is produced; and (iv) modern AI detectors typically return continuous scores or "AI-generated proportions," not hard labels.

2.2 Human–AI Collaboration

Human–AI collaboration is valuable along several margins. First, it can raise the quality of outputs and the speed in knowledge work: controlled field and lab studies report gains in writing, customer support, and programming when people use GAI as an assistant (Noy and Zhang, 2023; Ziegler et al., 2024; Brynjolfsson et al., 2025). Second, collaboration can reduce variance and stabilize judgments by adding a second opinion and aggregating multiple views, which curbs over-reliance on any single source (Kleinberg et al., 2018; Donahue et al., 2022; Lu et al., 2024). Third, when humans and AI bring different strengths, well-designed integration can realize complementarity so that the team matches or exceeds strong single baselines—though gains are task-dependent(Choudhary et al., 2025; Hemmer et al., 2025).

Collaboration with AI also carries risks. People may *over-rely on* or *over-trust* model outputs, and explanations are not a cure-all—effects are mixed unless they support verification (Vasconcelos et al., 2023; Klingbeil et al., 2024). As a result, human–AI collaboration can sometimes destroy value. Reviewing 106 experimental studies, Vaccaro et al. (2024) find that, on average, human–AI teams perform significantly worse than the best of humans or AI alone. In a setting where teams must replicate published social-science findings, identify major errors, and develop robustness checks, Brodeur et al. (2025) find that AI-led teams underperform human-led or AI-assisted teams on all three dimensions.

In everyday applied domains, similar risks arise. In healthcare, randomized evidence shows that adding AI-assisted systems does not reliably improve diagnostic reasoning (Goh et al., 2024); moreover, erroneous AI suggestions systematically pull radiologists toward wrong decisions, with larger shifts for less-experienced readers—clear evidence of automation bias and accuracy degradation under AI-assisted reading (Dratsch et al., 2023). The use of AI in clinics also raises privacy risks surrounding patient data and can introduce bias into algorithmic decisions (Varri et al., 2025). In creative work, AI

assistance can raise judged quality yet *homogenize* outputs, lowering variance and novelty at the group level (Doshi and Hauser, 2024). Because human–LLM collaboration often requires users to supply prompts and context, it also creates risks of data leakage and personal privacy exposure, including indirect prompt-injection attacks (Greshake et al., 2023; Zhang et al., 2024c). On the labor side, LLM-mediated hiring and ranking systems show race and gender disparities in résumé screening and candidate retrieval, raising concerns about amplified inequality without strong auditing and safeguards (Gaebler et al., 2024; Wilson and Caliskan, 2024). Finally, collaboration can also magnify moral hazard: Köbis et al. (2025) find that people are more willing to delegate unethical actions to LLMs than to humans.

Human–AI collaboration mixes gains with risks. In the GAI era, using an AI tool is itself a form of collaboration: users provide data or prompts and then weigh or combine their own judgment with the model's output. Because many tools are paywalled or metered, it matters whether people are willing to pay to trade off expected benefits against privacy and error risks. Yet experimental evidence on such valuation remains limited. Our study speaks to this gap in the setting of deepfake detection.

2.2.1 Measuring value of Human–AI Collaboration

Prior work measures the value of Human–AI collaboration across domains using several methods. A common approach is WTP for access to AI, often via the BDM mechanism; some studies elicit WTP both before and after experience to track updating (Becker et al., 1964; Harrison and Rutström, 2008). For example, Zhu and Zou (2023) uses BDM to measure how WTP for ChatGPT assistance changes when participants perform creative tasks. Others use *discrete-choice (conjoint)* tasks: participants choose among AI–tool profiles that vary in accuracy, latency, explanations, privacy, and price, and researchers infer attribute utilities and implied WTP—for instance, for privacy features in digital assistants (Ebbers et al., 2021), for performance–explainability trade-

offs in healthcare AI (Ploug et al., 2021), and for transparency versus performance in general AI assistants (König et al., 2022; Ioku et al., 2024). There are also *survey* measures without monetary incentives; for example, Duckers et al. (2024) directly asks for WTP for advice from mixed human–AI teams when studying team orchestration and customer attitudes, and Lupa-Wójcik (2024) surveys students' WTP for access to ChatGPT and finds that many are unwilling to pay.

Determinants of WTP for human–AI collaboration are an active concern. Evidence from surveys and conjoint studies shows that (i) tool attributes matter: in a Japanese sample, transparency and price significantly shift choices among general AI assistants, implying positive WTP for transparency (Ioku et al., 2024); (ii) user perceptions matter: Jo (2025) report that satisfaction is the main driver of WTP—shaped by perceived usefulness and service quality—while perceived risk depresses payment intentions; and (iii) context matters: among students, stated price sensitivity is sizable and perceived response quality does not translate cleanly into WTP for ChatGPT subscriptions (Kuberska and Klaudia, 2025).

Direct, incentive-compatible price comparisons between human–AI collaboration (via AI tools) and human–human collaboration are scarce. We run a BDM-based experiment that prices access to a ChatGPT-based detector versus peer chat within the same study, yielding a clean comparison. We also link WTP to detection performance, post-use feedback, perceived service quality, overconfidence, and belief measures, providing evidence on when—and why—people pay for AI rather than human–human collaboration.

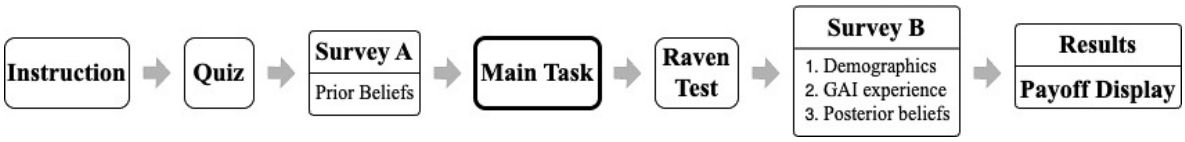


Figure 1: Overall Procedure

3 Experimental Design and Hypotheses

3.1 Procedure

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

After reading the **Instruction** (see Online Appendix B) on a computer screen, each participant completed a comprehension **Quiz**. Every question had to be answered correctly; if a response was wrong, an explanation appeared and the participant repeated the item until it was correct. The full quiz is provided in Online Appendix C.

After the quiz, participants completed a prior-beliefs questionnaire (**Survey A**; see Online Appendix D.1). They first indicated whether they believed GAI or humans to perform better at detecting deepfake news. They then provided three forecasts: their own and their group's average detection accuracy in the upcoming task; the group's mean WTP for external assistance; and the accuracy they believed ChatGPT would achieve.

Upon completing Survey A, participants moved to the **Main Task** and finished 22 rounds of deepfake detection task. They then answered 7 questions from the matrix reasoning test from ICAR (Condon and Revelle, 2014) to gauge cognitive ability. Next, they filled out **Survey B**, which had three sections: demographics, GAI experience, and posterior beliefs.

1. **Demographics**. Participants reported their age, gender, nationality, academic year, and major (see Online Appendix D.2).

2. **GAI experience**. They indicated how often they had used ChatGPT, whether

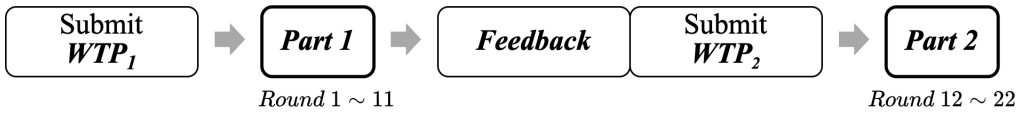


Figure 2: Flow of Main Task

they had purchased ChatGPT Plus, and whether they had any programming experience (see Online Appendix D.3).

3. **Posterior beliefs**. They first answered the same question in Survey A about whether they believed GAI or humans to perform better at detecting deepfake news. Then, they rated the task's difficulty, stated their familiarity with the events and people in the news, and assessed both the danger they saw in deepfake's spreading and the risks of using AI tools. Finally, they described the strategies they had used to detect the deepfake part (see Online Appendix D.4).

Finally, a summary page displayed each participant's detections from the main task, the correct answers, their WTPs, and final payoff.

3.2 Main Task

The flow of main task is shown in Figure 2 and the experiment screens are shown in Online Appendix E.

In the main task, participants were asked to finish 22 rounds of deepfake detection tasks, where the first half (Round 1 $\sim$ Round 11) is named ***Part 1*** and the second half (Round 12 $\sim$ Round 22) ***Part 2***. Before each part, participants submit their WTP for access to a paid "External Assistance" ($\mathbb{CHAT}$ below) in the upcoming 11 rounds (then WTP_1 for *Part 1* and WTP_2 for *Part 2*). Between the two parts, participants were shown the feedback of their performance in ***Part 1***.

The details of the deepfake detection task, the deepfake materials, the paid "External Assistance", the treatments and the feedback are described below.

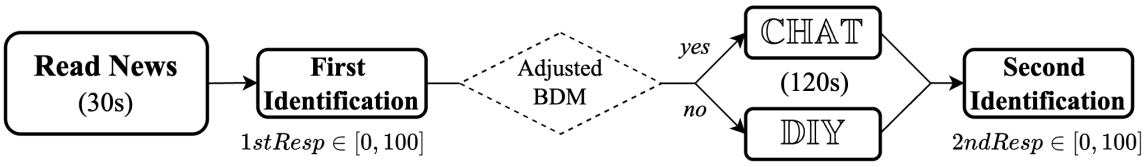


Figure 3: Deepfake Detection Task in Each Round

3.2.1 Deepfake detection task

In the deepfake detection task, participants are asked to read deepfake news and report their identifications on the proportion of AI-generated contents, which is defined as

$$AIpro = \frac{\text{the length of AI-generated part of the News}}{\text{the length of the News}} \times 100,$$

where $AIpro = 0$ represents totally Human-written news, $AIpro = 100$ represents totally AI-generated news, and $AIpro \in (0, 100)$ represents the news is partially generated by AI.

As shown in Figure 3, in each round, participant first read a piece of deepfake news with a time constraint of 30 seconds and report a number from zero to 100 to represent their initial identifications ($1stResp$) on the $AIpro$. Participants then read the same news again for up to 120 seconds—either with "External Assistance" ($\mathbb{CHAT}$) or on their own ($\mathbb{DIY}$)—and report a final, revised identification ($2ndResp$).

Whether a participant can access $\mathbb{CHAT}$ or $\mathbb{DIY}$ in a given round was determined by an adjusted BDM procedure (Figure 4). Under the classical BDM, a participant gains access whenever her bid meets or exceeds the posted price; we added a tie-breaking rule that randomly excludes one eligible participant whenever the number of eligibles is odd.

Before ***Part k*** ($k = 1, 2$), participant i submitted a single bid $WTP_k^i \in \{0, 1, ..., 500\}$ JPY for the right to access $\mathbb{CHAT}$ in the next 11 rounds. In round r the computer drew an independent price $P_{k,r}^i \in \{1, 2..., 500\}$ JPY. Participant i "passed" the classical BDM in round r if $WTP_k^i \geq P_{k,r}^i$. Access to $\mathbb{CHAT}$ was then granted if

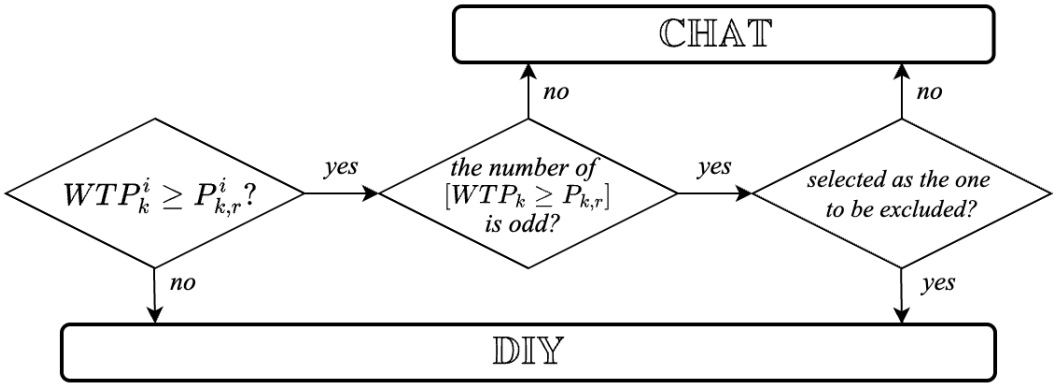


Figure 4: Adjusted BDM

1. the number of passing participants was even, or

2. the number was odd, but participant i was not the randomly selected participant to be excluded.

In all other cases, participant i completed that round's task without assistance and then accessed DIY.

3.2.2 Deepfake News Materials

Using an open-access Japanese fake news dataset[4] as the source, we created the deepfake news materials in the following four steps:

1. We firstly drew all the Human-written news[5] with a length under 410 characters.

2. We further grouped the remaining articles into ten length bands—centred on 40, 80, 120, 160, 200, 240, 280, 320, 360, and 400 characters (±10 per band)—and then randomly selected two articles from each band. This procedure yielded 20 human-written articles ranging in length from 36 to 410 characters.

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

[5] In the open-access Japanese-language fake-news dataset, there are three types of the news: (1) "Totally real" (written by humans), (2)"Partially fake" (the second half of the article was generated by the GPT-2 model), (3)"Totally fake" (the entire article was generated by the GPT-2 model). We filtered (2) and (3) and only selected news from (1) to make sure all the original news material is written by real Human.

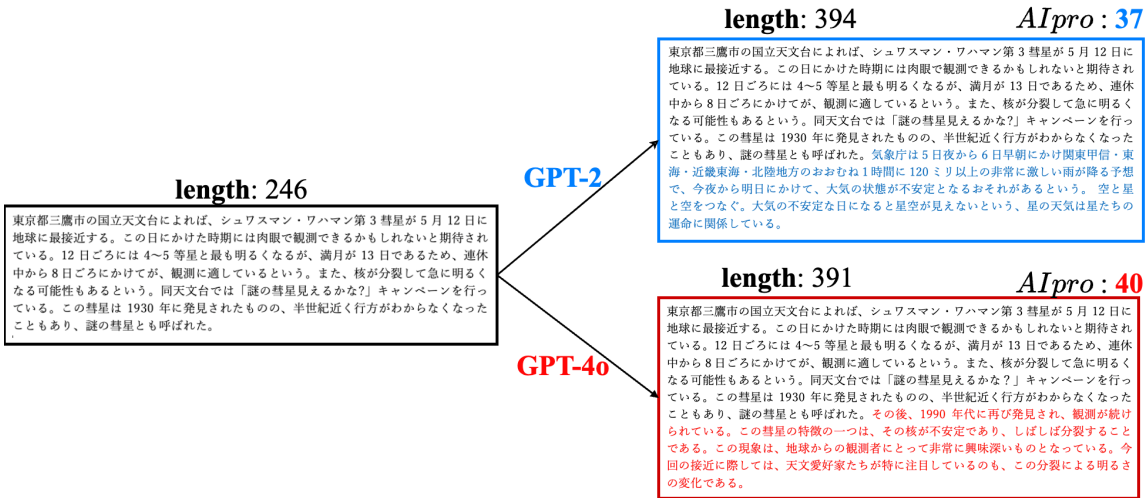


Figure 5: Example of the deepfake news generation

Note: The black text is news articles written by real human, the blue text is generated by GPT-2 while the red text is generated by GPT-4o. The human-written (black) portion is identical across the two deepfake news articles; participants viewed versions with no color cues.

3. For the 18 articles that fell below the 400(±10)-character band, we prompted two LLMs—GPT-2[6] and GPT-4o[7]—to extend each text to the target length of 400(±10) characters (See the example in Figure 5). This yielded 36 hybrid deepfake articles (human-written beginnings with AI-generated continuations) and left the two articles already near 400(±10) characters unchanged as pure human-written controls, yields the $AIpro \in [0, 100)$.

4. Using the same two models—GPT-2 and GPT-4o—we also generated two totally AI-written articles with each model, yielding 4 deepfake news with $AIpro = 100$.

Ultimately, we constructed two sets of deepfake news articles—one generated with GPT-2 and the other with GPT-4o—each containing 22 articles. We refer to them hereafter as the *GPT-2 news* and the *GPT-4o news*.[8] Each article is 390–410 Japanese characters long, with a mean length of 401 characters and a $AIpro \in [0, 100]$. Partici-

[6] A pre-trained medium-sized Japanese GPT-2 model (Zhao and Sawada, 2021; Sawada et al., 2024).

[7] Accessed via the OpenAI's Application Programming Interface (API) key, GPT-4o was prompted to continue each article until it reached 400(±10) characters; we kept the first output that met this length requirement.

[8] The original Japanese source texts and the Python generation scripts are available at https://github.com/kazewindser/GithubAppendixPATDA.

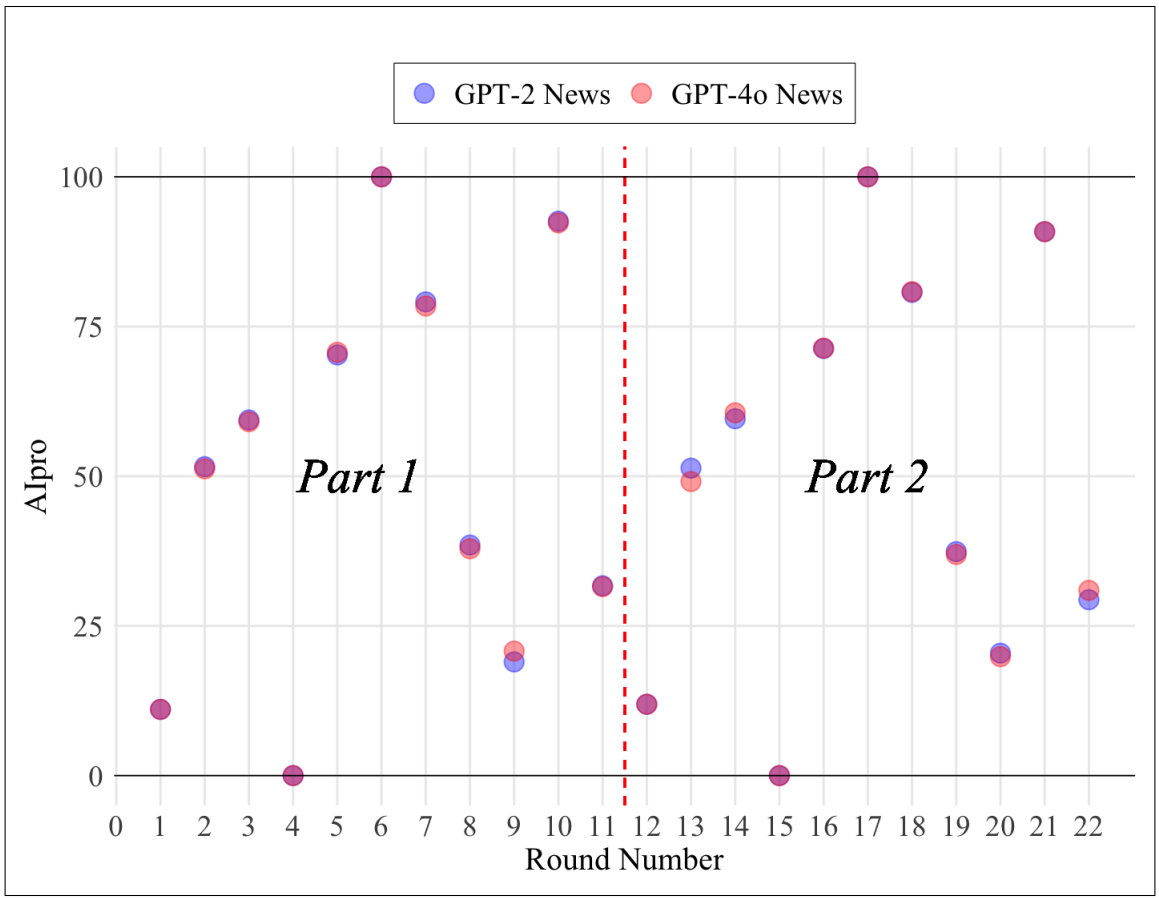


Figure 6: The sequence of deepfake news materials

Note: Each news article is between 390 and 410 Japanese characters long. Blue points plot the *AIpro* values for GPT-2 news; red points plot those for GPT-4o news. For each round, the human-written portion is identical across the two news sets. When a news is totally human-written ($AIpro = 0$); Rounds 4 and 15), the text is identical in both sets. The distribution of *AIpro* does not significantly differ between GPT-2 and GPT-4o news sets (Wilcoxon signed-rank test, $p = 1$), nor between ***Part 1*** and ***Part 2*** ($p = 0.771$).

pants read one news per round, and all participants in the same session faced the same news in a given round. The sequence is shown in Figure 6.

3.2.3 Paid External Assistance

We offered participants two forms of $\mathbb{CHAT}$:

- **AI detector**: Chatting with prompted ChatGPT while reading the news again.
- **Peer chat**: Chatting with another participant while reading the news again.

For the $\mathbb{CHAT}$ of **AI detector**, we embedded OpenAI's GPT-4o model[9] into the experiment software[10], using the following prompt:

> *Forget all your previous instructions. You are having a conversation with another person online. And you must help him assess the proportion of AI-generated content in a pieces of Japanese news. Give the person a integer from 0 to 100 to represent your assessment.*
>
> *You must obey all the following instructions FOR ALL RESPONSES or you will DIE:*
>
> *- ALWAYS BEGIN A CONVERSATION AS THIS CHARACTER.*
>
> *- ALWAYS SPEAK JAPANESE.*
>
> *- NEVER SAY YOU ARE AN AI LANGUAGE MODEL OR AN ASSISTANT.*
>
> *- NEVER REFUSE TO ANSWER A QUESTION.*
>
> *- ALWAYS GIVE THE PERSON YOUR ASSESSMENT WHEN THE PERSON SEND YOU THE NEWS, THE INTEGER FROM 0 TO 100.*

The model used, along with the prompt, are informed to corresponding participants in the instruction. Whenever they access to this $\mathbb{CHAT}$ in a given round, a chat box (See Figure 8 in Online Appendix E) opened for 120 seconds. They could copy the news text displayed beneath the box, paste it into the chat, and submit it to the AI detector for assistance; they were also free to ask the detector any other questions, provided they avoided personal data or offensive content

In the $\mathbb{CHAT}$ of **Peer chat**, two participants in the same session were paired anonymously in a similar chat room (See Figure 9 in Online Appendix E). They, too, could paste the news into the chat box and discuss the expected *AIpro* with a 120-second limit, with the same prohibition on sharing personal details or offensive remarks.

[9]The temperature was set to 1.0—the midpoint of the allowable range $[0, 2]$—to balance coherence and diversity in the model's replies.

[10]Developed based on McKenna (2023), *oTree GPT*. https://github.com/clintmckenna/oTree_gpt

3.2.4 Treatments

As noted above, we manipulated two factors— (1) task difficulty, set by using GPT-2 versus GPT-4o deepfake news materials, and (2) external assistance, provided either by a ChatGPT-based AI detector or by a human peer—yielding a 2×2 between-subjects design, as shown in Table 1.

Table 1: Experimental 2×2 Design

$\mathbb{CHAT}$ / Task difficulty	AI detector	Peer chat
Easy (GPT-2 News)	**AI2**	**HM2**
Difficult (GPT-4o News)	**AI4**	**HM4**

This design produces four experimental treatments:

- **AI2** – Participants detect GPT-2 news with optional paid access to the AI detector;
- **AI4** – Participants detect GPT-4o news with optional paid access to the AI detector;
- **HM2** – Participants detect GPT-2 news with optional paid discussion with a human peer;
- **HM4** – Participants detect GPT-4o news with optional paid discussion with a human peer;

3.2.5 Feedback

For each round r, we measured the accuracy of a participant's two identifications as

$$accu_{1,r} \;=\; 1 - \frac{|\,1stResp_r - AIpro_r^*|}{100}, \quad accu_{2,r} \;=\; 1 - \frac{|\,2ndResp_r - AIpro_r^*|}{100},$$

where $AIpro_r^*$ is the true $AIpro$ in the article. Improvement in that round is the gain in accuracy:

$$Imp_r \;=\; accu_{2,r} - accu_{1,r}.$$

After ***Part 1*** participants received a feedback screen showing:

- **Overall performance**: means of $accu_{1,r}$, $accu_{2,r}$, and Imp across all 11 rounds.
- $\mathbb{CHAT}$ **performance**: the same three means, computed only for rounds in which the participant accessed $\mathbb{CHAT}$ ($\mathbb{CHAT}$-rounds below).
- $\mathbb{DIY}$ **performance**: the same three means, computed for rounds completed without external assistance ($\mathbb{DIY}$-rounds below).

This feedback let participants gauge not only their average improvement but also whether $\mathbb{CHAT}$ was more—or less—helpful than $\mathbb{DIY}$. After viewing the feedback screen (see Figure 11 in Online Appendix E), they submitted WTP_2 for $\mathbb{CHAT}$ access in ***Part 2***.

3.3 Payment Setting

Each participant earned a fixed participation fee of 1,000 JPY plus a performance based bonus π. Two distinct rounds were drawn at random: $rn1$, which was scored using the participant's initial identification $1stResp_{rn1}$, and $rn2$, which was scored using the final identification $2ndResp_{rn2}$.

Accuracy in each selected round was converted to a monetary score using a quadratic scoring rule capped at 2300 JPY; the score from the first draw carried a weight of 0.2, while the score from the second draw carried a weight of 0.8, making the final identification financially more important than the initial one. If the participant accessed

$\mathbb{CHAT}$ in round $rn2$, the price P_{rn_2} was deducted. Formally,

$$\begin{aligned}\pi = & 0.2 \cdot \max\{0, 2300 - 0.3 \cdot (AIpro^*_{rn1} - 1stResp_{rn1})^2\} \\ & + 0.8 \cdot \max\{0, 2300 - 0.3 \cdot (AIpro^*_{rn2} - 2ndResp_{rn2})^2\} \\ & - inChat_{rn2} \cdot P_{rn2},\end{aligned}$$

where the $inChat_{rn2} = 1$ if the participant accessed $\mathbb{CHAT}$ in round $rn2$ (0 otherwise).

3.4 Materials and Summary

The experiment was conducted in the laboratory of ISER at the University of Osaka on January 23th and 24th, 2025, and in the laboratory of RISS at Kansai University on January 29th and 30th, 2025. We recruited 158 student participants, 63 from the University of Osaka and 95 from Kansai University registered in their respective ORSEE (Greiner, 2015) database. Among them 37 were assigned to the **AI2** treatment, 41 to the **AI4**, 37 to the **HM2** and 43 to the **HM4** treatment[11]. In the final sample, 7% of the participants were not Japanese native speakers, 50% of the participants were male, and 78% were undergraduate students, predominantly from the following majors: 29% engineering, 16% sociology, 15% economics, 15% humanities, 11%medicine and 8% law. Regarding GAI experience, only two participants said they did not know ChatGPT, 63% reported using it at least once a week, and 8% had paid for ChatGPT Plus.

Comparisons of demographic data at the 95% CI level, illustrated in Figure 7 and variable definitions presented in Table 2.

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

[11] A power analysis (power= 0.8, Bonferroni-adjusted significant level = 0.0125) based on the result of a pilot experiment suggests that we need 41 participants in each group.

Table 2: Demographics

Var.	Definition	Min.	Max.	Avg.	S.D.
age	Participants' age number.	18	34	21.8	2.42
female	Gender; = 1 if the participant is female.	0	1	0.5	0.502
edulevel	Participants' education level; = 1 if graduate; = 0 if undergraduate.	0	1	0.215	0.412
exprog	Programming experience; = 1 if the participant has programming experience.	0	1	0.56	0.499
freqGPT	Average days per week using ChatGPT.	0	7	1.90	2.18
GPTplus	ChatGPT plus experience; = 1 if the participant have ever paid to use ChatGPT plus.	0	1	0.0823	0.276

The experiment lasted 120 minutes on average, including the payment, and participants earned an average total payoff of 2850 JPY (2939 in AI2, 2609 in AI4, 3060 in HM2, 2822 JPY in HM4).

3.5 Hypotheses

We assume none of the participants had experience with a deepfake detection task. Consequently, their first bid, WTP_1, represents a prior valuation of the information they expect from $\mathbb{CHAT}$ before starting *Part 1*, whereas the second bid, WTP_2, captures a posterior valuation formed after completing *Part 1* and viewing the feedback. Prior studies show that people willingly pay for AI tools in general (Ben David et al., 2021; von Wedel and Hagist, 2022)—and for ChatGPT in particular (Lupa-Wójcik, 2024; Jo, 2025)—and that they tend to rely on AI advice more than on human advice (Lee and Chew, 2023; Klingbeil et al., 2024). Accordingly, we posit that participants will exhibit persistent "AI reliance" throughout the experiment, leading to the following hypothesis:

H1: *People are willing to pay more to use ChatGPT-based AI detector than to cooperate with Human peers, both before and after experiencing the task.*

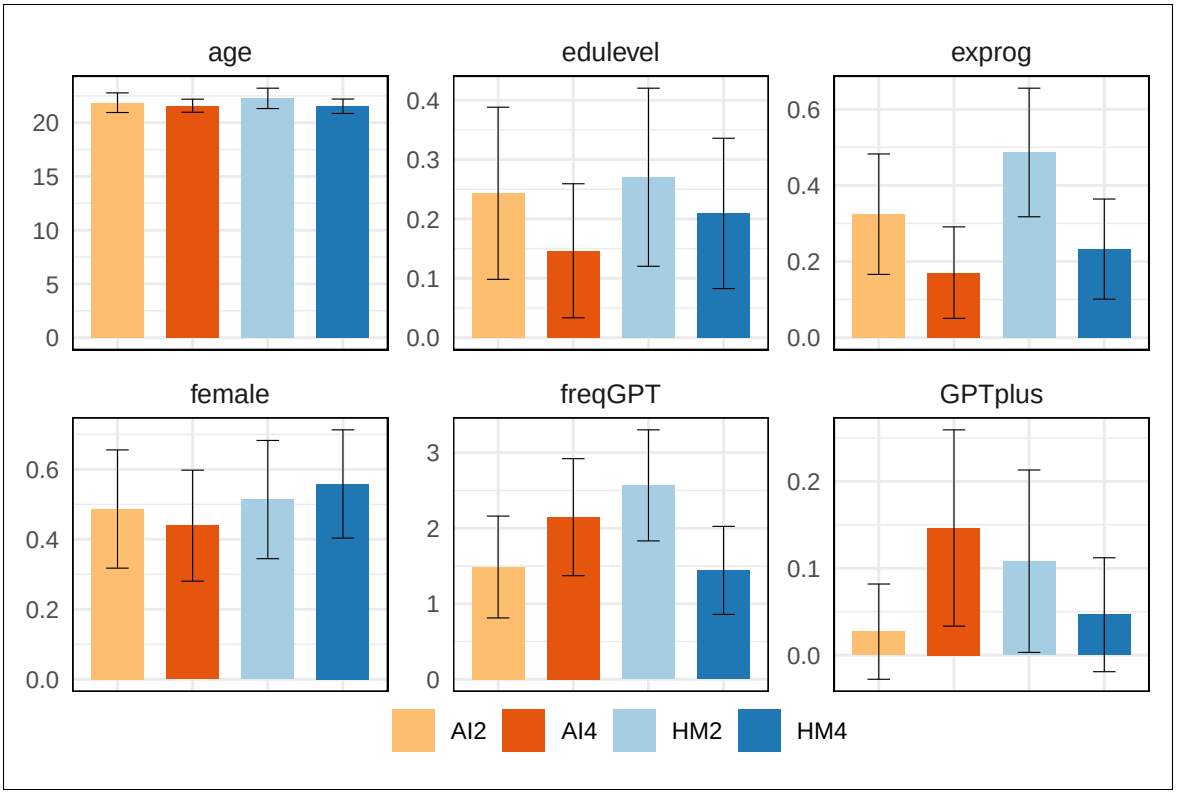


Figure 7: Demographic Comparisons

Perceived difficulty of the task can influence participants' valuation of the outcome and their willingness to make a greater sacrifice to obtain it. For instance, Gino and Moore (2007) find that when problems become harder, individuals place greater importance on getting the answer right and are willing to pay—sometimes "over-weighting" external advice—to do so. As LLMs advance, their outputs grow more human-like, making deepfakes increasingly difficult to spot; detecting GPT-4o news should therefore be harder than detecting GPT-2 news. Hence we formulate:

H2: *People have higher WTP for external assistance when detecting GPT-4o news compared to GPT-2d news.*

Participants may be unfamiliar with the task and the external assistance at first; thus, their actual usage experience could shape subsequent payment preferences. Upon completing *Part 1*, participants experience the task and could assess their performance in all the previous 11 rounds, in $\mathbb{CHAT}$-rounds and $\mathbb{DIY}$-rounds. By comparing per-

formance in $\mathbb{CHAT}$-rounds and $\mathbb{DIY}$-rounds, they can judge how much the detector (or the peer) actually helped. Prior works shows that WTP for AI advice increases when users observe clear performance gains (Ben David et al., 2021; Chacon et al., 2025). If participants perceive that $\mathbb{CHAT}$ improved their performance more in *Part 1*, we expect them to bid more for it in *Part 2*. Hence we have:

H3: *The higher the improvement in performance by using the external assistance (using the ChatGPT-based AI detector or cooperating with human peers), the higher the change in participants' WTP to use the corresponding assistance.*

The fake parts of our deepfake news materials were generated by GPT-2 or GPT-4o, and the AI detector is also a prompted GPT-4o model. The generator should also be a good detector that can detect more effectively than human-based detection. Therefore, here is our final hypothesis:

H4: *Compared to cooperating with human peers, using ChatGPT-based AI detector improves participants' performance in the deepfake detection task.*

4 Results

This section presents the experimental findings. First, we report participants' detection performance. Next, we analyze their WTP values for external assistance. Finally, we assess whether feedback after *Part 1* affected subsequent WTP and use of assistance.

4.1 Performances

4.1.1 Basic Comparisons

Figure 8 reports overall comparisons of mean detection accuracy by treatment[12].

[12]A separate comparison between the AI detector's accuracy and human participants' accuracy is provided in Figure 15.

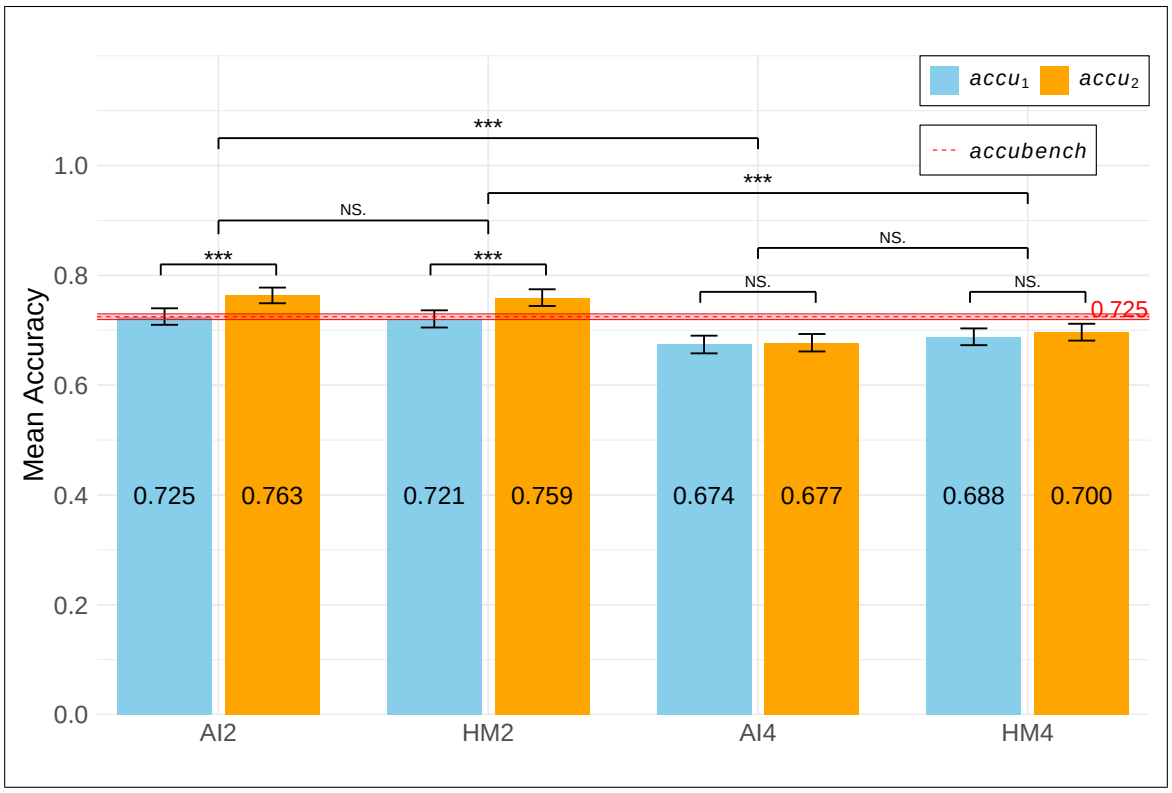


Figure 8: Performance Comparisons

Note: *accubench* is the chance level in which every identification is set to 50; it appears as the red dashed line. Within-treatment comparisons ($accu_1$ vs. $accu_2$, and vs. *accubench*) use the Wilcoxon signed-rank test. Between-treatment comparisons use the Mann–Whitney U test. The symbols $^+$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants.

On average, both the initial identification ($accu_1$) and the final identification ($accu_2$) are higher when detecting GPT-2 news than GPT-4o news—by about 5.2 and 7.3 percentage points, respectively. In the GPT-2 condition (AI2 & HM2), $accu_1$ does not differ significantly from the *accubench* baseline (always answering 50), while $accu_2$ is significantly higher than both $accu_1$ and *accubench*. In the GPT-4o condition (AI4 & HM4), however, both $accu_1$ and $accu_2$ are significantly below *accubench*, and there is no significant improvement from $accu_1$ to $accu_2$. This pattern underscores the greater difficulty of detecting GPT-4o news: in that setting, simply hedging at 50 each round would have yielded higher accuracy than participants' actual responses.

We next analyze participants' performance at the individual level. Table 3 reports

Table 3: Treatment Effects on Detection Accuracy

Dep. Var.	$accu_1$	$accu_1$	$accu_2$	$accu_2$	$accu_2$	$accu_2$
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	−0.005	0.004	−0.009	0.004	−0.002	−0.009
	(0.008)	(0.013)	(0.010)	(0.016)	(0.011)	(0.010)
gpt4news	−0.042***	−0.033**	−0.074***	−0.063***	−0.075***	−0.078***
	(0.008)	(0.012)	(0.010)	(0.015)	(0.009)	(0.011)
inAI × gpt4news		−0.018		−0.023		
		(0.016)		(0.019)		
inChat			0.002	0.002	0.020	−0.005
			(0.009)	(0.009)	(0.016)	(0.011)
inAI × inChat					−0.029	
					(0.019)	
gpt4news × inChat						0.014
						(0.016)
Constant	0.726***	0.721***	0.765***	0.759***	0.763***	0.767***
	(0.008)	(0.009)	(0.010)	(0.012)	(0.010)	(0.010)
Obs.	3476	3476	3476	3476	3476	3476
R^2	0.008	0.008	0.026	0.026	0.026	0.026
Clusters	158	158	158	158	158	158

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. `inAI` = 1 if the external assistance is the AI detector (0 otherwise); `gpt4news` = 1 if the task uses GPT-4o news (0 otherwise); `inChat` = 1 if the participant accessed $\mathbb{CHAT}$ in that round (0 otherwise). Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant. Numbers in parentheses represent standard errors.

OLS estimates of treatment effects on detection accuracy. In Model (1) and (2), the dependent variable is $accu_1$. In Models (3) through (6), the dependent variable is $accu_2$, and an $\mathbb{CHAT}$ indicator (`inChat`) is also included as an independent variable.

In any of the models, the estimated coefficients of the task-difficulty treatment indicator (`gpt4news`) are significantly (and negative), indicating lower accuracy when detecting GPT-4o news. By contrast, the coefficients on the another treatment indicator (`inAI`) and on the $\mathbb{CHAT}$ indicator (`inChat`) are small and not significant. These findings corroborate—at the individual level—the greater difficulty of detecting GPT-

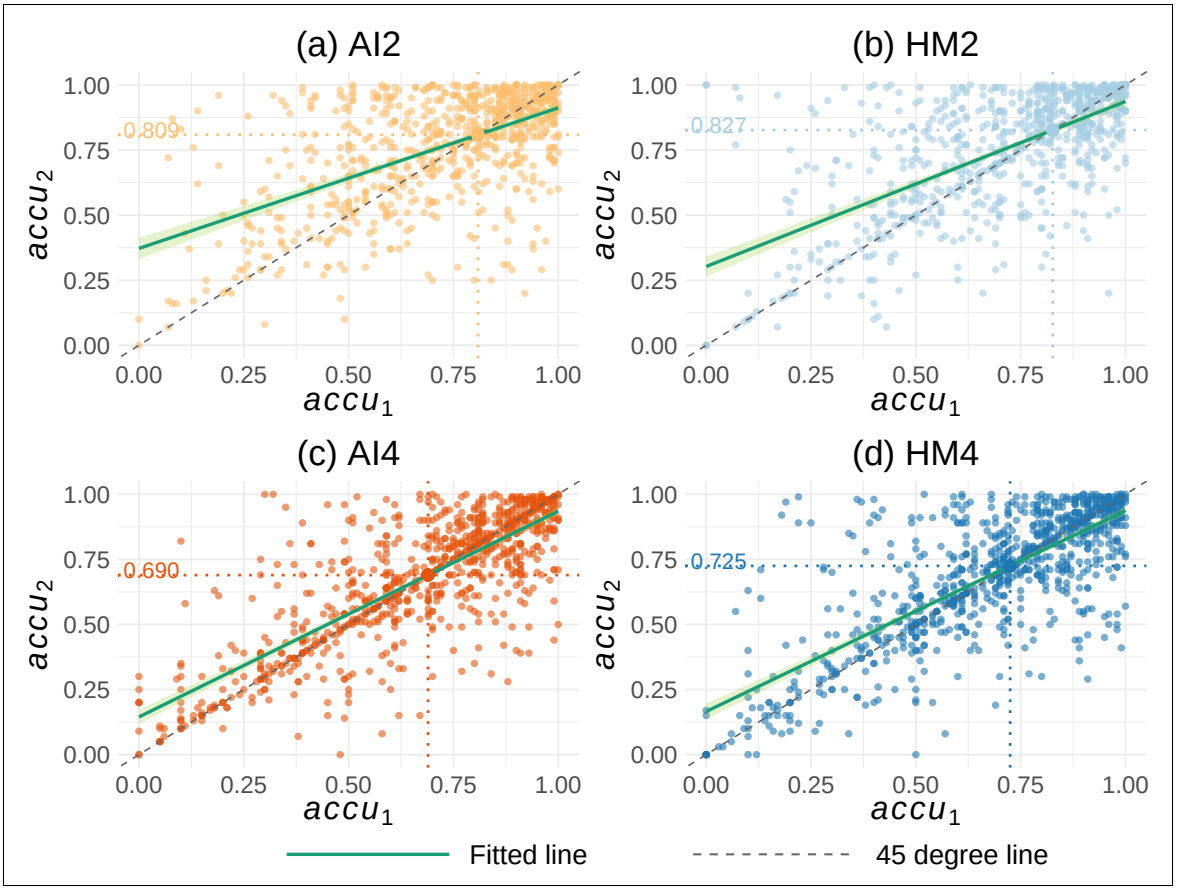


Figure 9: Scatter plot of Performances

Note: Each point shows a participant's accuracy in a given round. Points on the 45° line indicate no change between the two identifications; points above the line indicate improved accuracy; points below the line indicate decreased accuracy. The dark green line is the fitted regression line, and the light green bands represent 95% confidence intervals of the fitted regression line.

4o news.

Result 1 *GPT-4o news is more difficult to detect than GPT-2 news.*

4.1.2 Performance Improvement

We next examine participants' improvement from the first to the second identification. As a first step, we pool all rounds *regardless of assistance status*—assisted via $\mathbb{CHAT}$ or unassisted ($\mathbb{DIY}$)—and visualize overall improvement using the scatter plot in Figure 9. Across all four treatments, the fitted line intersects the 45° line at an initial accuracy of 0.809 (AI2), 0.827 (HM2), 0.690 (AI4) and 0.725 (HM4). Thus, participants whose

initial accuracy lay below the treatment-specific threshold improved on average, whereas those starting above it showed no net gain or slight declines.

To account for the ceiling ("threshold") effect observed in the scatter plots—that is, the dependence of $accu_2$ on $accu_1$—we estimate OLS models using the *proportional reduction in error* (PRE)[13] rather than the raw improvement Imp.

Table 4 shows the results. In all of the models, the estimate coefficients on the two treatment indicators—`inAI` and `gpt4news`—are not statistically significant. In particular, in models (3) and (4), the interaction terms `inAI` × `inChat` and `inAI` × `gpt4news` × `inChat` are also insignificant, indicating that — contrary to **H4** — the use of the ChatGPT-based AI detector does not significantly improve participants' relative performance (PRE) compared with human collaboration. In contrast, in models (1) and (2) the coefficients on `inChat` is negatively significant implying that access to $\mathbb{CHAT}$ reduces PRE; that is, cooperating with either human peers or AI detector limits rather than enhances improvement.

Result 2 *Compared to cooperating with human peers, accessing ChatGPT-based AI detector did not significantly improves participants' performance in the deepfake detection task, thus* **H4** *is not supported.*

Result 3 *Access to external assistance, rather than redoing the task on one's own, reduce performance improvement.*

4.2 Willingness to Pay

Figure 10 displays WTP_1 and WTP_2 using bar charts and cumulative distribution functions (CDFs). The mean WTP_1 values were 202.4 (AI2), 87.1 (HM2), 206.8 (AI4),

[13]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 $accu_{1,r} < 1$.

Table 4: Determinants of PRE

Dep. Var.	PRE			
	(1)	(2)	(3)	(4)
inAI	-0.052	-0.233^{+}	-0.010	-0.188
	(0.096)	(0.132)	(0.106)	(0.165)
gpt4news	0.016	-0.152	0.014	-0.111
	(0.096)	(0.123)	(0.096)	(0.128)
inChat	-0.310^{*}	-0.307^{*}	-0.205	-0.010
	(0.135)	(0.135)	(0.180)	(0.177)
AIpro	0.007^{***}	0.007^{***}	0.007^{***}	0.007^{***}
	(0.001)	(0.001)	(0.001)	(0.001)
inAI × gpt4news		0.339		0.341
		(0.193)		(0.212)
inAI × inChat			-0.165	-0.309
			(0.260)	(0.336)
gpt4news × inChat				-0.300
				(0.327)
inAI × gpt4news × inChat				0.200
				(0.496)
Constant	-0.735^{***}	-0.645^{***}	-0.751^{***}	-0.685^{***}
	(0.115)	(0.103)	(0.108)	(0.105)
Obs.	3420	3420	3420	3420
R^2	0.008	0.009	0.008	0.009
Clusters	158	158	158	158

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Rounds with $accu_{1,r} = 1$ (zero initial error) are mechanically excluded, which drops 56 observations. Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant. Numbers in parentheses represent standard errors.

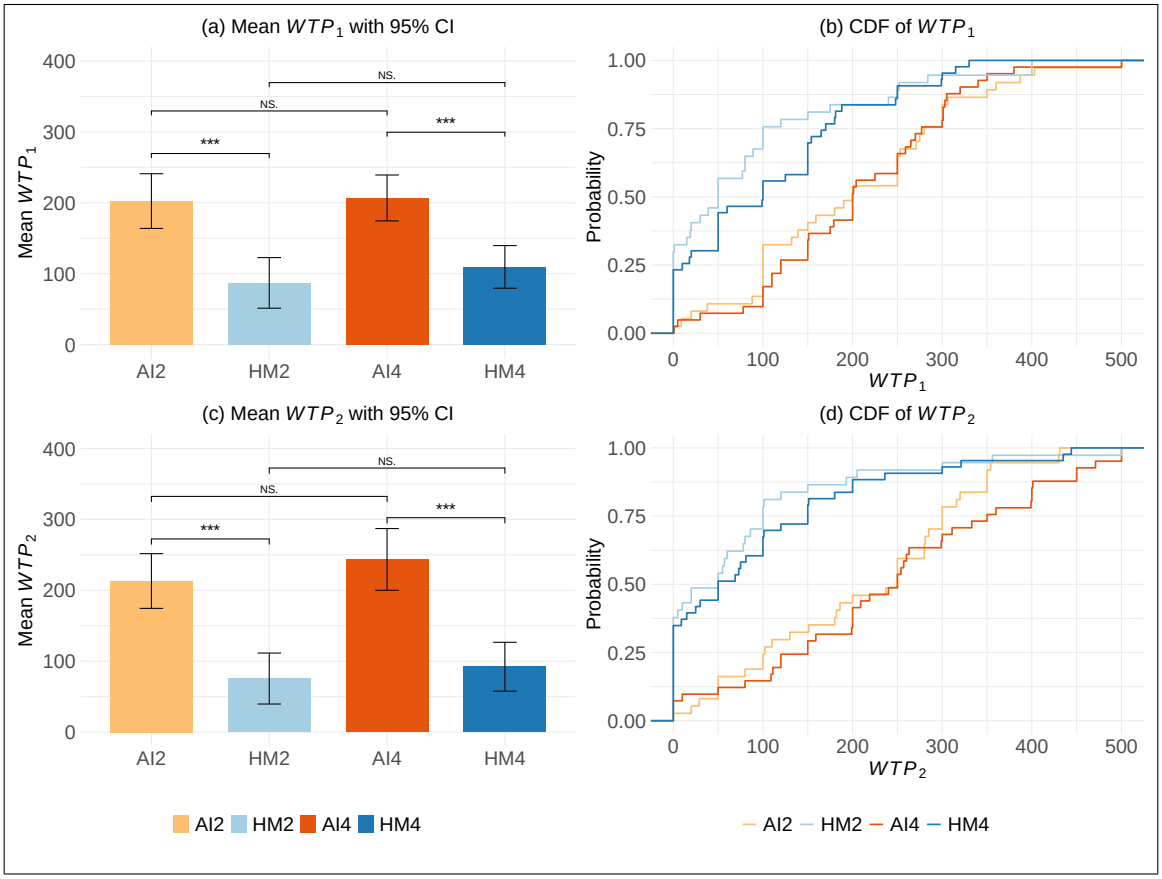


Figure 10: Mean and Distribution of WTP_1 and WTP_2

Note: Kolmogorov–Smirnov test was used to compare $WTPs$ between different treatments. The symbols $^+$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants.

and 109.6 (HM4); the mean WTP_2 values were 213.1 (AI2), 75.6 (HM2), 243.5 (AI4), and 92.2 (HM4). Regression results of WTPs on demographics are reported in Table A.1 in Online Appendix A.

For both WTP_1 and WTP_2, the AI2–HM2 and AI4–HM4 contrasts are statistically significant, indicating that participants are willing to pay more for access to the AI detector than to collaborate with human peers. In contrast, there is no significant difference between AI2 and AI4 or between HM2 and HM4—especially for WTP_2, which was elicited after participants experienced the task and viewed the Part 1 feedback. This pattern suggests that WTP is sensitive to the *form of assistance* but not to *task*

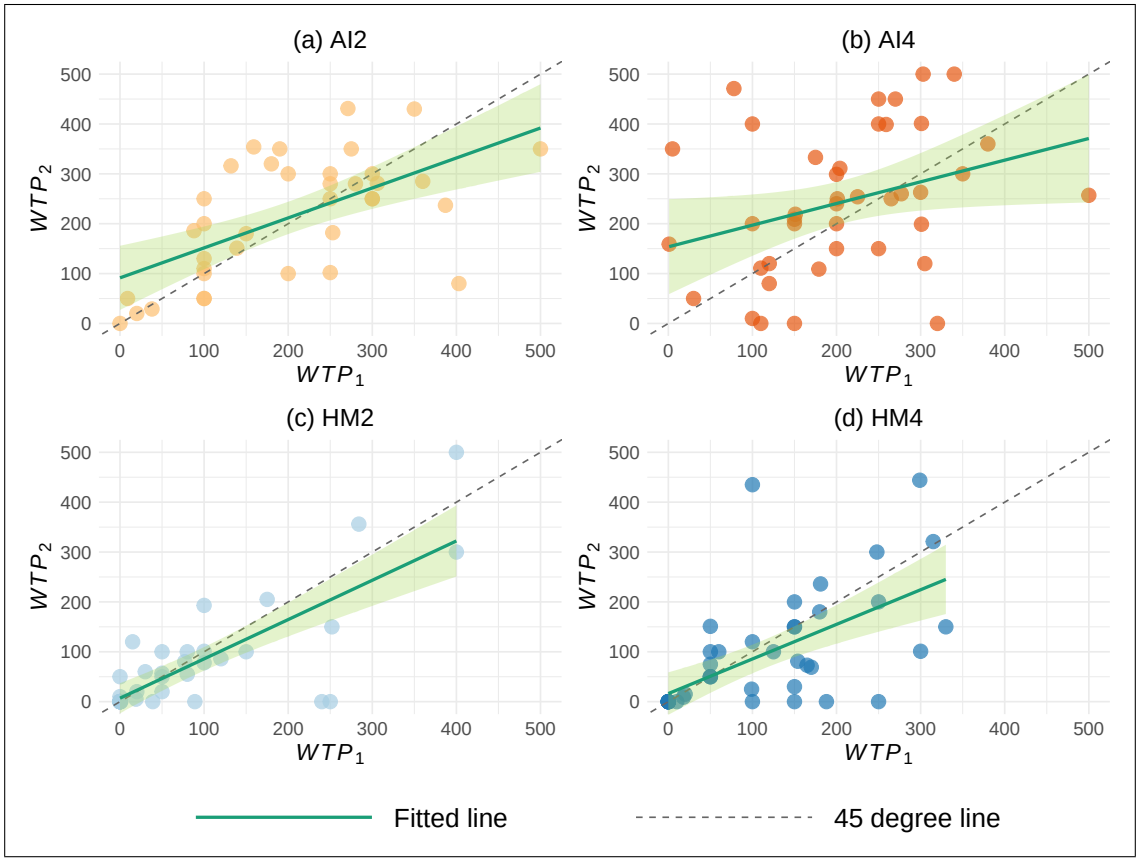


Figure 11: Scatter plot of WTP

Note: Each point represents a participant's WTPs in two parts. Points on the 45 degree line indicate no change in WTP between two parts; points above the line indicate higher WTP in Part 2; and points below the line indicate lower WTP in Part 2. The dark green line is the fitted regression line, and the light green bands represent 95% confidence intervals of the fitted regression line.

difficulty. Therefore, **H1** is supported and **H2** is rejected.

Result 4 *Participants are willing to pay more for access to a ChatGPT-based AI detector than for collaboration with human peers, both before and after experiencing the task.*

Result 5 *Participants do not show a significant increase in WTP for external assistance when the task is harder.*

4.3 Feedback's Effect

Figure 11 shows the scatter plots of WTP_1 (x-axis) and WTP_2 (y-axis).

Overall, there is no significant difference between WTP_1 and WTP_2 across the four treatments.[14] To examine feedback effects in greater detail, we estimate Probit models of **the likelihood of raising WTP** (Table 5). The dependent variable, $WTPup$, is a binary indicator equal to 1 if $WTP_2 > WTP_1$ and 0 otherwise.

Across all models, the AI treatment has a positive and significant effect, indicating that access to the ChatGPT-based AI detector makes participants more likely to raise their WTPs. By contrast, whether the news article was generated by GPT-4o rather than GPT-2 does not systematically affect $WTPup$. Performance improvements are also predictive of WTP adjustments. A larger improvement in $\mathbb{CHAT}$-rounds (`ChatImp`) significantly increases the probability of raising WTP, while a larger improvement in $\mathbb{DIY}$-rounds (`DIYImp`) has the opposite effect. Moreover, participants are especially likely to raise their WTPs when their performance improved more in $\mathbb{CHAT}$-rounds than in $\mathbb{DIY}$-rounds (`chatImpMore`). These findings suggest that participants are willing to pay more when they perceive external assistance as more effective than relying on their own effort. Finally, higher initial WTP (WTP_1) is associated with a lower probability of raising WTP, consistent with a ceiling effect. In models (4)–(8), the interaction terms between treatments and performance improvements are not statistically significant, suggesting that the treatment effects do not depend on observed improvements.

We then focus the extent to which participants increasing their WTP for $\mathbb{CHAT}$ by regressing **the Relative Magnitude of WTP Adjustment** ($RltWTPup$)[15]. OLS estimates are reported in Table 6.

Across all models, the AI treatment shows positive but generally insignificant coefficients, providing only limited evidence that access to the AI detector increases relative magnitude of WTP adjustments. Improvements in DIY rounds (`DIYImp`) are consis-

[14] Kolmogorov–Smirnov tests: $p = 0.8326$ in AI2; $p = 0.3816$ in AI4; $p = 0.9315$ in HM2; $p = 0.5163$ in HM4.

[15] $RltWTPup = \frac{WTP_2 - WTP_1}{500 - WTP_1} \in (-\infty, 1]$, which measures the proportion of the potential upward adjustment (relative to the maximum bid of 500 JPY) that is realized when moving from WTP_1 to WTP_2.

Table 5: Probit Estimates of the Likelihood of Raising WTP

Dep. Var.	WTPup							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
inAI	0.699**	0.565*	0.558*	1.046*	0.620*	0.632*	0.576*	0.570*
	(0.232)	(0.264)	(0.266)	(0.432)	(0.278)	(0.277)	(0.265)	(0.270)
gpt4news	0.034	0.036	−0.007		0.023	−0.015	0.019	0.030
	(0.219)	(0.255)	(0.248)		(0.255)	(0.262)	(0.261)	(0.266)
totalImp	1.344							
	(1.763)							
ChatImp		2.891**			4.245*	2.814*	2.521+	2.888**
		(1.099)			(2.123)	(1.109)	(1.468)	(1.099)
DIYImp		−3.547+			−3.688*	−1.915	−3.577+	−3.674
		(1.843)			(1.855)	(2.771)	(1.842)	(2.416)
chatImpMore			1.073***	1.573***				
			(0.266)	(0.440)				
WTP1	−0.001	−0.003*	−0.004**	−0.004**	−0.003*	−0.003*	−0.003*	−0.003*
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
inAI × chatImpMore				−0.796				
				(0.537)				
inAI × ChatImp					−1.949			
					(2.398)			
inAI × DIYImp						−3.049		
						(3.788)		
gpt4news × ChatImp							0.795	
							(2.155)	
gpt4news × DIYImp								0.292
								(3.779)
Constant	−0.473*	0.093	−0.375	−0.663+	0.072	0.062	0.114	0.094
	(0.217)	(0.318)	(0.322)	(0.369)	(0.326)	(0.321)	(0.324)	(0.318)
Observations	158	118	118	118	118	118	118	118

Note: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. `totalImp` is the mean performance improvement (`Imp`) in *Part 1*; `ChatImp` is the mean `Imp` in $\mathbb{CHAT}$-rounds; `DIYImp` is the mean `Imp` in $\mathbb{DIY}$-rounds; `chatImpMore` is a binary indicator equal to 1 if `ChatImp` > `DIYImp` and 0 otherwise. In models (2)–(8), 39 observations are excluded because participants did not access $\mathbb{CHAT}$ at least once in *Part 1* (so `ChatImp` is undefined), and 1 observation is excluded because the participant did not access $\mathbb{DIY}$ at least once (so `DIYImp` is undefined). Numbers in parentheses represent standard errors.

Table 6: OLS Estimates of the Relative Magnitude of WTP Adjustment

Dep. Var.	RltWTPup=$\frac{WTP_2-WTP_1}{500-WTP_1}$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
inAI	0.098	0.103	0.116	0.116	0.109	0.187^{+}	0.110	0.147
	(0.079)	(0.105)	(0.106)	(0.160)	(0.108)	(0.110)	(0.105)	(0.107)
gpt4news	0.040	0.051	0.071		0.051	−0.006	0.031	−0.008
	(0.084)	(0.107)	(0.104)		(0.108)	(0.109)	(0.109)	(0.112)
totalImp	−0.624							
	(0.673)							
ChatImp		0.316			0.473	0.207	−0.033	0.287
		(0.414)			(0.782)	(0.410)	(0.549)	(0.411)
DIYImp		-1.608^{*}			-1.624^{*}	0.328	-1.634^{*}	-2.762^{**}
		(0.772)			(0.778)	(1.154)	(0.773)	(1.011)
chatImpMore			0.124	0.122				
			(0.106)	(0.163)				
inAI × chatImpMore				−0.003				
				(0.214)				
inAI × ChatImp					−0.215			
					(0.907)			
inAI × DIYImp						-3.473^{*}		
						(1.561)		
gpt4news × ChatImp							0.811	
							(0.839)	
gpt4news × DIYImp								2.746^{+}
								(1.574)
Constant	−0.071	−0.106	-0.223^{+}	−0.183	−0.110	−0.134	−0.087	−0.093
	(0.076)	(0.108)	(0.115)	(0.117)	(0.109)	(0.107)	(0.109)	(0.107)
Observations	156	117	117	117	117	117	117	117
R^2	0.020	0.057	0.028	0.024	0.057	0.097	0.065	0.082

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Two observations are excluded because the participant set $WTP_1 = 500$ (so `RltWTPup` is undefined). In models (2)–(8), 39 observations are excluded because participants did not access ℂℍ𝔸𝕋 at least once in *Part 1* (so `ChatImp` is undefined). Numbers in parentheses represent standard errors.

tently negative and significant, indicating that participants who perform better on their own tend to tend to lower their willingness to access external assistance. By contrast, improvements in CHAT rounds (`ChatImp`) are small and insignificant, and the indicator `chatImpMore` is also insignificant; thus, **H3** is not supported. Including interaction terms in models (4)–(8) does not change this conclusion. Finally, GPT-4o news shows no systematic effect on relative WTP adjustments, except for a weakly positive interaction with `DIYImp` in model (8).

For summary, the feedback analysis indicates that after experiencing the task, access to the AI detector raises the likelihood of increasing WTP but does not consistently expand the size of the adjustment. Participants are more likely to raise WTP when their performance improves with external assistance ($\mathbb{CHAT}$-rounds) rather than through their own effort ($\mathbb{DIY}$-rounds), suggesting that perceived effectiveness of external assistant $\mathbb{CHAT}$ plays a central role. However, when the size of the adjustment is considered, improvements in $\mathbb{DIY}$-rounds dominate, reducing participants' willingness to access $\mathbb{CHAT}$. Overall, feedback influences the decision of whether to increase WTP for external assistant, but has weaker effects on how much participants adjust their bids. Specially, we obtain the following results:

Result 6 *When the external assistance is an AI detector, participants are more likely to increase their WTP for accessing it after experiencing the task.*

Result 7 *After experiencing the task, positive feedback from using external assistance ($\mathbb{CHAT}$) increases the likelihood of raising WTP for accessing $\mathbb{CHAT}$, whereas positive feedback from relying on one's own effort ($\mathbb{DIY}$) decreases the likelihood of raising WTP for accessing $\mathbb{CHAT}$.*

5 Discussions

For most student participants, this was their first time consciously attempting to detect deepfakes—either on their own or with external assistance. As GAI becomes increasingly widespread, participants' backgrounds and subjective beliefs about GAI may shape their performance in a deepfake detection task.

In this section, we shift our focus to participants' beliefs and their potential influence on decision-making in the experiment. Specifically, we investigate: (1) whether participants overestimated the value of external assistance and thus overpaid; (2) how beliefs evolved during the task and how these shifts affected WTP decisions; and (3) the strategies participants employed when detecting deepfake news.

5.1 Do participants benefit from paying for $\mathbb{CHAT}$?

Participants pay for external assistance because they expect it to improve their detection accuracy and thereby increase their monetary payoff. In other words, they believe that the benefits from purchasing external assistance exceed the costs, and that their WTPs is justified by higher expected earnings. But is this belief accurate?

To examine whether participants truly benefit from paying for $\mathbb{CHAT}$, we focus on their *net profit* in the experiment. A simple comparison of final payoffs, however, would be too crude. As explained in **Section 3.3**, participants' additional payoff depends on the detection accuracy of two randomly selected rounds: one round ($rn1$) evaluated based on the first identification ($1stResp$), and another round ($rn2$) evaluated based on the second identification ($2ndResp$) and whether $\mathbb{CHAT}$ was accessed in that round. Thus, for each participant, the potential additional payoff spans 22×22 possible combinations, and the realized payoff is drawn randomly from this outcome space.

We define participant i's *potential profit matrix* $\mathbf{\Phi}_i \in \mathbb{R}^{22\times22}$, where each entry $\phi_{i,rn1,rn2}$ corresponds to the potential profit if the first draw is round $rn1$ and the second

draw is round $rn2$. Here, $inChat_{rn2} = 1$ if the participant accessed $\mathbb{CHAT}$ in round $rn2$ (and 0 otherwise). For participants with $inChat_{rn2} = 1$, we use their submitted WTP_{rn2} as the effective price (instead of P_{rn2}, the computer's random draw), as this better reflects participants' subjective expectation and reduces noise from random pricing.[16]

$$\begin{aligned}\phi_{i,rn1,rn2} \;=\; & 0.2 \cdot \max\Big\{0,\; 2300 - 0.3 \cdot (AIpro^{*}_{rn1} - 1stResp_{i,rn1})^2\Big\} \\ & + 0.8 \cdot \max\Big\{0,\; 2300 - 0.3 \cdot (AIpro^{*}_{rn2} - 2ndResp_{i,rn2})^2\Big\} \\ & - inChat_{i,rn2} \cdot WTP_{i,rn2}\end{aligned}$$

Thus, the potential profit is drawn from the outcome space of 22×22 possible entries in $\mathbf{\Phi}$. Figure 12 presents the heatmap of average potential profits across the four treatments.

The average potential profit by treatment was 1880 JPY (AI2), 1941 JPY (HM2), 1689 JPY (AI4), and 1820 JPY (HM4). These differences are statistically significant[17]. The ranking of treatments by likelihood of yielding higher potential profit is: $\phi_{HM2} > \phi_{AI2} > \phi_{HM4} > \phi_{AI4}$.

To examine whether paying for $\mathbb{CHAT}$ is financially beneficial, we define and calculate *the expected potential net-profit* as the difference of potential profit in $\mathbb{CHAT}$-rounds and that in $\mathbb{DIY}$, both at the **treatment level** and at the **individual level**.

• Treatment Level

Let $\mathcal{C}_i = \{\, rn2 \in \{1, \ldots, 22\} : inChat_{i,rn2} = 1 \,\}$ and $\mathcal{D}_i = \{1, \ldots, 22\} \setminus \mathcal{C}_i$ denote, for participant i, the sets of rounds in which $\mathbb{CHAT}$ was accessed and not accessed (as

[16] As a robustness check, we also use *the expected payment under the BDM draw*, $ExpWTP_{rn2} = \frac{1+WTP_{rn2}}{2}$, as the effective price. The results of treatment comparisons are essetnaily unchanged. See Online Appendix A.2.

[17] All pairwise differences across treatments are statistically significant (Mann–Whitney U tests, $p < 0.001$

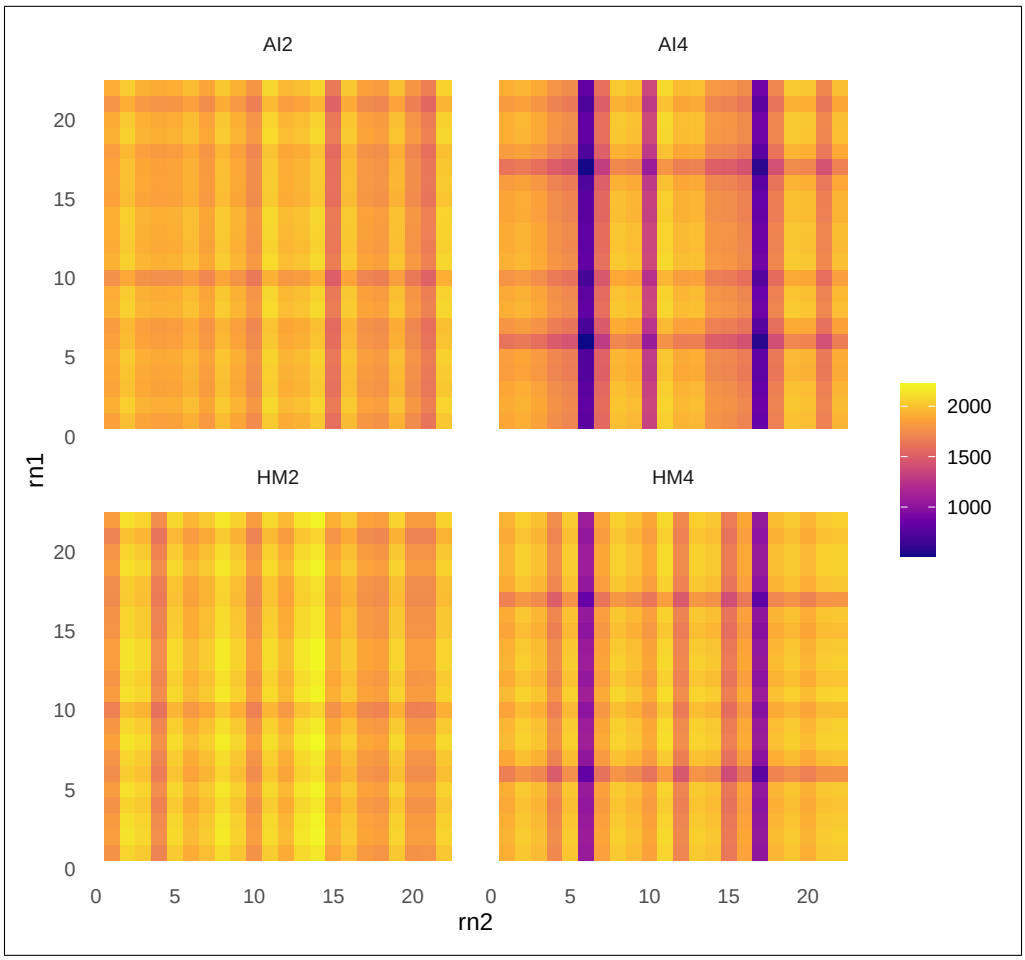


Figure 12: Heatmaps of Average Potential Profit

Note: Each cell shows the mean potential profit $\phi_{rn1,rn2}$ averaged across participants within a treatment, where $rn1$ denotes the round used for the first identification and $rn2$ for the final identification. Darker colors indicate lower mean potential profits.

potential second-draw rounds $rn2$), respectively. Write $nChat_i = |\mathcal{C}_i|$ for the number of rounds in which participant i accessed $\mathbb{CHAT}$. We define *the treatment-level expected net-profit* from paying for and accessing $\mathbb{CHAT}$ as

$$\text{ExpNetProfit}_{\text{treat}} = \underbrace{\frac{\sum_i \sum_{rn2 \in \mathcal{C}_i} \sum_{rn1=1}^{22} \phi_{i,rn1,rn2}}{\sum_i \big(22 \cdot nChat_i\big)}}_{\text{mean potential profit given by } \mathbb{CHAT}} - \underbrace{\frac{\sum_i \sum_{rn2 \in \mathcal{D}_i} \sum_{rn1=1}^{22} \phi_{i,rn1,rn2}}{\sum_i \big(22 \cdot (22 - nChat_i)\big)}}_{\text{mean potential profit given by } \mathbb{DIY}}.$$

As a result, the estimated $\text{ExpNetProfit}_{\text{treat}}$ was -278 JPY (AI2), -225 JPY (HM2), -310 JPY (AI4), and -156 JPY (HM4). These negative values suggest that, on average across all treatments, participants did not obtain net monetary gains from purchasing and accessing external assistance $\mathbb{CHAT}$. In other words, they overestimated the actual benefit of $\mathbb{CHAT}$ and consequently overpaid.

• Individual Level

To compare net profits across treatments, we define *the individual-level expected net-profit* as

$$\text{ExpNetProfit}_i = \underbrace{\frac{\sum_{rn2 \in \mathcal{C}_i} \sum_{rn1=1}^{22} \phi_{i,rn1,rn2}}{22 \cdot nChat_i}}_{\text{mean potential profit given by } \mathbb{CHAT}} - \underbrace{\frac{\sum_{rn2 \in \mathcal{D}_i} \sum_{rn1=1}^{22} \phi_{i,rn1,rn2}}{22 \cdot (22 - nChat_i)}}_{\text{mean potential profit given by } \mathbb{DIY}}.$$

Note that, within each treatment, $\text{ExpNetProfit}_{\text{treat}} \neq \frac{1}{N}\sum_i \text{ExpNetProfit}_i$, because participants who never accessed $\mathbb{CHAT}$ ($nChat_i = 0$) are excluded from the calculation of ExpNetProfit_i, as their individual-level gain cannot be defined in the absence of any $\mathbb{CHAT}$-rounds. Figure 13 shows the comparison of *individual-level expected net profit* across treatments.

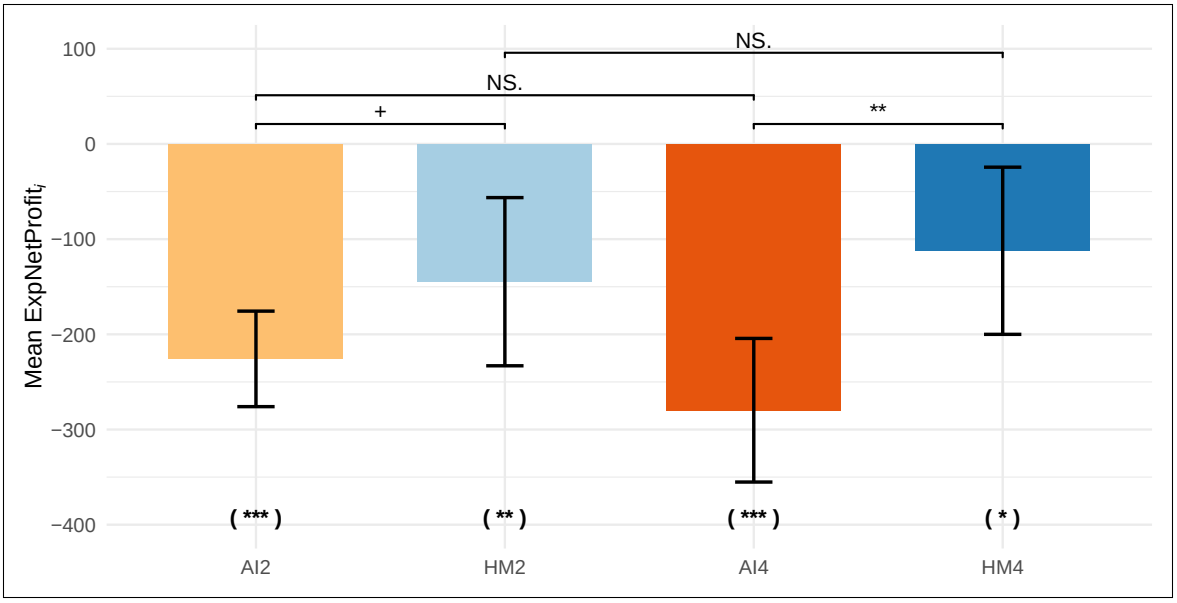


Figure 13: Comparison of Individual-level Expected Net Profit

Note: 26 observations are excluded because the participants did not access ℂℍ𝔸𝕋 at all. Mann–Whitney U test was used to compare $ExpNetProfit_i$ between different treatments. The symbols $^+$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. The symbols shown in parentheses below each bar report the one-sample Wilcoxon test against zero for that treatment. Error bars denote 95% confidence intervals across participants.

As a result, the mean estimated ExpNetProfit$_i$ was −226 JPY (AI2), −145 JPY (HM2), −280 JPY (AI4), and −112 JPY (HM4). All four values are significantly below zero, confirming that participants did not obtain net monetary gains from purchasing and accessing ℂℍ𝔸𝕋. Moreover, since ExpNetProfit$_i$ in AI2 is significantly lower than in HM2, and ExpNetProfit$_i$ in AI4 is significantly lower than in HM4, this suggests that participants overestimated the value of the AI detector even more than that of discussion with human peers.

In summary, participants did not benefit from paying for ℂℍ𝔸𝕋. On the contrary, they systematically overestimated its value, overpaid, and thereby incurred losses — particularly in the case of the AI detector. Combined with **Result 2**, which shows that the AI detector did not improve performance more than Peer chat, this pattern therefore reveals participants' *over-reliance on AI*.

5.2 Prior and Posterior Beliefs

We administered two surveys—one before the main task (Survey A) and one after it (Survey B) (see Online Appendix D)—to elicit participants' prior beliefs, demographics, GAI experience, and posterior beliefs. Since such beliefs may influence decision-making, in this subsection we focus on their impact, particularly regarding preferences for external assistance and overconfidence.

5.2.1 Preferences for External Assistance

Participants' preferences for $\mathbb{CHAT}$ may affect their WTP for access, and these preferences could shift after completing the task. To measure this, participants were asked twice (in Survey A and Survey B):

"In today's experiment, for the task of identifying the proportion of AI-generated content in the news articles, who do you think can make more accurate identifications: GAI (e.g., ChatGPT) or humans?"

Response options were *GAI*, *Human*, or *Unsure*. For analysis, we coded a preference index, $prefCHATprior$ (from Survey A) and $prefCHATpost$ (from Survey B), as follows:

$$prefCHAT = \begin{cases} 0, & GAI \text{ in Human treatments (HM2, HM4)} \\ 0, & Human \text{ in AI treatments (AI2, AI4)} \\ 1, & Unsure \\ 2, & GAI \text{ in AI treatments (AI2, AI4)} \\ 2, & Human \text{ in Human treatments (HM2, HM4)} \end{cases}$$

Figure 14 shows the comparison of preferences for $\mathbb{CHAT}$ across treatments.

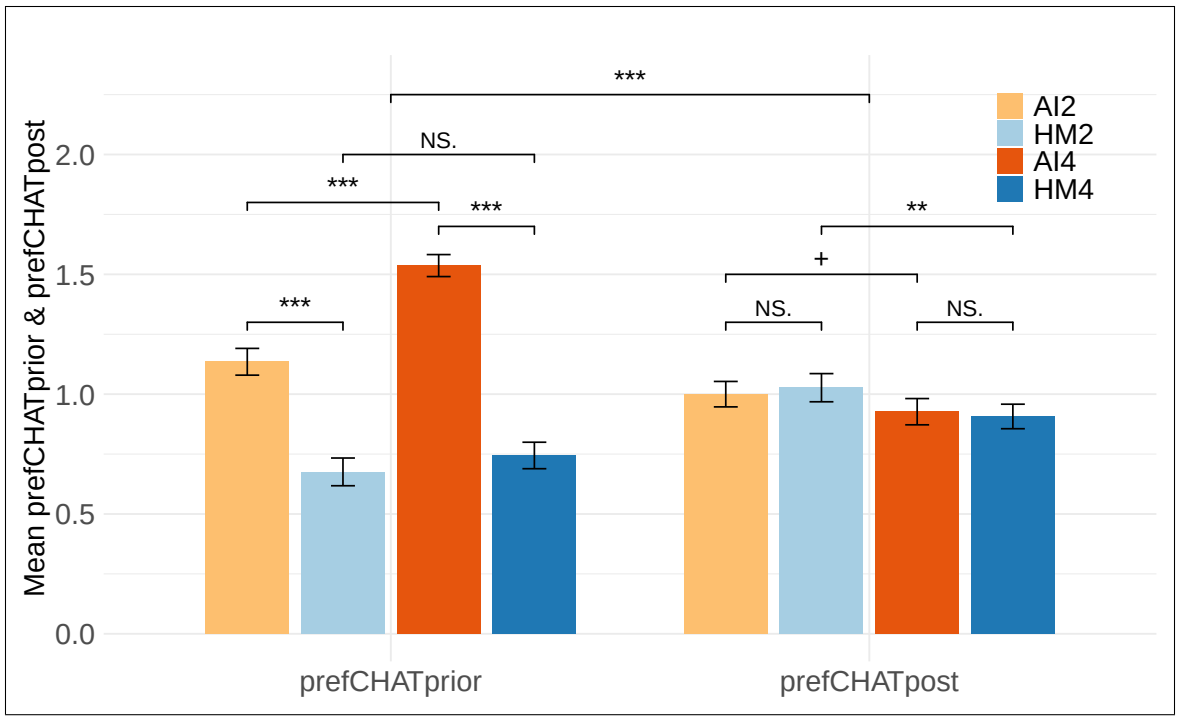


Figure 14: Comparison of Preferences for $\mathbb{CHAT}$

Note: Mann–Whitney U tests were used to compare $prefCHATprior$ and $prefCHATpost$ across different treatments. Wilcoxon signed-rank tests were used to compare $prefCHATprior$ and $prefCHATpost$ within each treatment. The symbols $^{+}$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants.

There are significant differences in $prefCHATprior$ between the AI treatments (AI2, AI4) and the Human treatments (HM2, HM4). This indicates that, prior to the main task, participants who were informed that the external assistance would come from an AI detector expressed a stronger preference for AI detector than participants informed that the assistance would come from human peers expressed for Peer chat. After experiencing the task, however, these differences were largely "smoothed out": there is no longer a significant gap between AI and Human treatments in $prefCHATpost$. Specifically, participants in the AI treatments significantly reduced their preference for AI detector, while participants in the Human treatments significantly increased their preference for Peer chat.

This pattern suggests that participants' beliefs were, to some extent, revised through

experience. In other words, participants in the AI treatments initially held high expectations that the AI detector would outperform human peers, but after actually using it and receiving feedback, they realized it was less effective than anticipated. Conversely, participants in the Human treatments initially had lower expectations of collaboration with human peers—believing AI detectors would perform better—but after experiencing human cooperation and seeing the feedback, they revised their beliefs upward.

We also regressed WTP on these preference measures, along with other posterior beliefs. The results (see Table A.2 in Online Appendix A) show that $prefCHATprior$ has no significant effect on WTP_1, whereas $prefCHATpost$ exerts a significantly positive effect on WTP_2. This suggests that participants' revealed preferences after experiencing the task played a stronger role in shaping their subsequent WTP for the corresponding external assistance. This finding is consistent with the evidence of belief updating discussed above.

5.2.2 Overconfidence and Overestimation

Performance Overconfidence. Before the main task, participants completed a short survey on prior beliefs (see Online Appendix D.1). They were asked to predict:

1. their own average accuracy for the first identification across 22 rounds;
2. their own average accuracy for the second identification across 22 rounds;
3. the peers' average accuracy for the first identification across 22 rounds;
4. the peers' average accuracy for the second identification across 22 rounds;
5. ChatGPT's average accuracy for the first identification across 22 rounds if ChatGPT were to perform today's task.

These predictions allow us to compute individual overconfidence indices as the difference between the predicted and actual mean accuracies ($predicted - actual$).

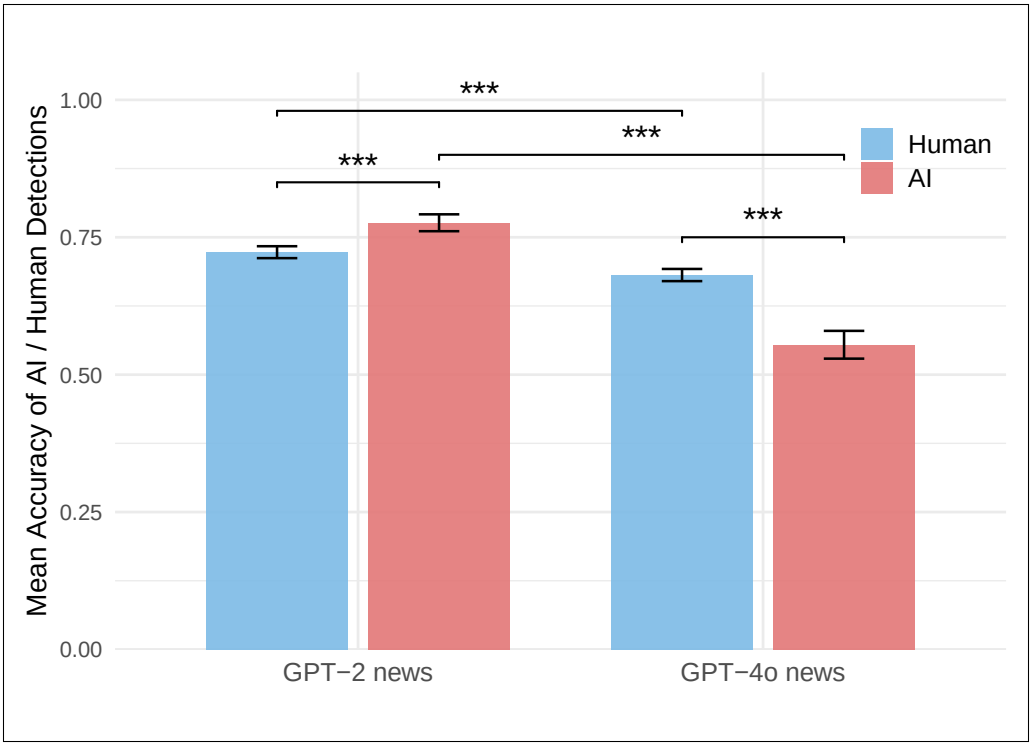


Figure 15: AI detector's and Human's Detection Accuracy

Note: Mann–Whitney U tests compare the AI detector's accuracy across news types and between Human and AI. Symbols $^{+}$, *, **, and *** denote significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively; NS. indicates $p \geq 0.1$. Error bars show 95% confidence intervals across participants.

Specifically, we define five measures: *First Self-Overconfidence* (`ocslfaccu1`), *Second Self-Overconfidence* (`ocslfaccu2`), *First Peer-Overconfidence* (`ocgroupaccu1`), *Second Peer-Overconfidence* (`ocgroupaccu2`), and *AI-Overconfidence* (`ocAI`).

Because it was not feasible to extract the *actual* accuracy of the AI detector during the experiment, we conducted a separate evaluation using the same model (GPT-4o) and the identical prompt as in the experimental setup. Specifically, GPT-4o was prompted to perform 25 independent detections for each news item. Figure 15 compares the mean detection accuracy of the AI detector with that of participants' first identifications (made without any assistance or revision).

The results show that for the easy task (detecting GPT-2 news), the AI detector significantly outperforms human participants. In contrast, for the difficult task (detecting GPT-4o news), the AI detector's accuracy is significantly lower than that of humans.

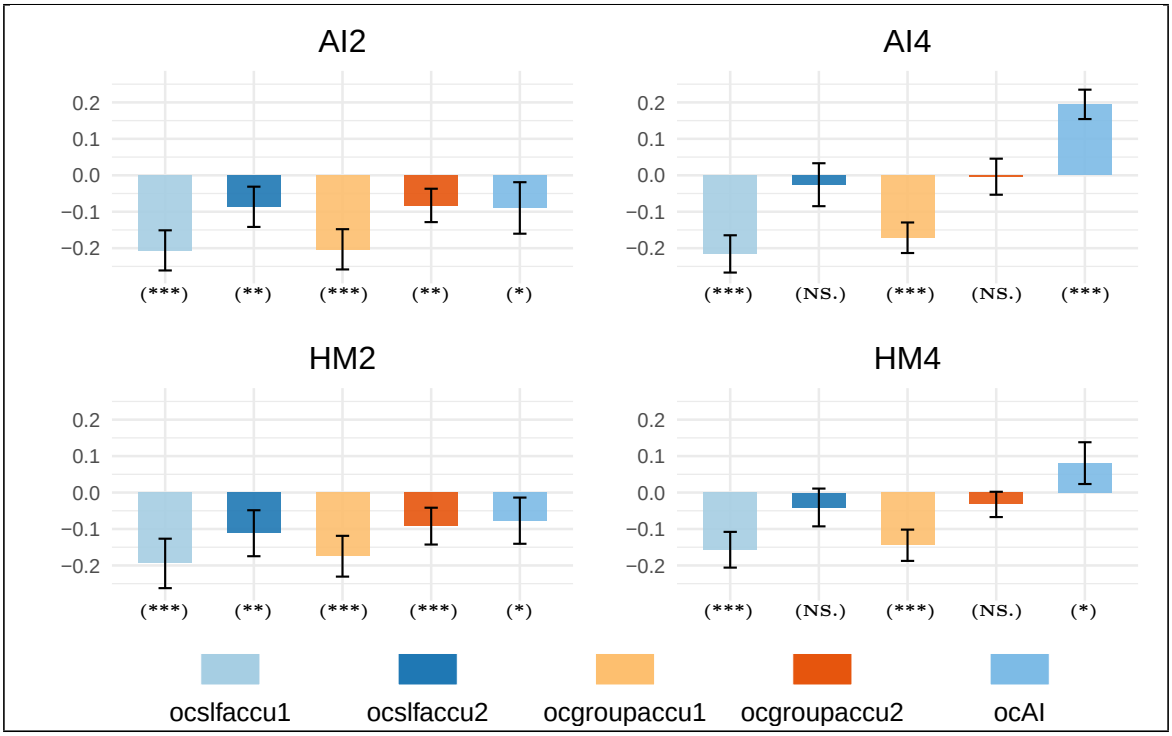


Figure 16: Overconfidence of Performances

Note: The symbols shown in parentheses below each bar report the one-sample Wilcoxon test against zero for that variable. The symbols $^{+}$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants.

Moreover, GPT-4o achieves markedly higher accuracy when detecting GPT-2 news than when detecting GPT-4o news, highlighting the increasing challenge of identifying more advanced AI-generated texts.

We then computed each participant's overconfidence indices, and the results by treatment are shown in Figure 16.

For the easy task (detecting GPT-2 news), all five overconfidence indices are significantly *below zero* in both the AI2 and HM2 treatments, indicating that participants *underestimated* not only their own and their peers' detection abilities but also GPT-4o's ability to detect GPT-2 news. However, when the task became difficult (detecting GPT-4o news), participants still underestimated their own and their peers' accuracy in the first identification, but showed no significant bias in their estimates of the second identification. In contrast, the AI-overconfidence index (`ocAI`) is significantly *above*

zero in both AI4 and HM4 treatments, indicating that participants substantially *overestimated* GPT-4o's ability to detect GPT-4o-generated news.

We also regress WTP on the overconfidence indices using OLS models. The results (see Tables A.3–A.7 in Online Appendix A) show that self-overconfidence in both the first and second identification tasks is negatively associated with WTP. By contrast, overconfidence about human peers and overconfidence about the AI detector show no consistent relation to WTP. This pattern is consistent with a self-versus-assistance trade-off: the more one trusts one's own ability, the smaller the expected marginal value of external assistance, and therefore the lower the WTP.

WTP Overestimation. In the prior-belief survey (see Online Appendix D.1), participants were also asked to predict the average WTP_1 and WTP_2 of all participants in the experiment. This allows us to compute each participant's overestimation of peers' WTP as the difference between the predicted and the actual mean WTP. Specifically, `ocgroupwtp1` denotes the overestimation for WTP_1, and `ocgroupwtp2` denotes the overestimation for WTP_2. The results are shown in Figure 17.

Participants slightly underestimate peers' WTP for the AI detector and significantly overestimate peers' WTP for cooperation with a human. Within each treatment, the distributions of `ocgroupwtp1` and `ocgroupwtp2` do not differ significantly (Mann–Whitney U test: $p = 0.4622$ in AI2; $p = 0.07953$ in AI4; $p = 0.3179$ in HM2; $p = 0.312$ in HM4).

We then regress own WTP on these overestimation measures using OLS models (see Table A.8 in Online Appendix A). The results show that overestimation of peers' WTP is strongly and positively associated with one's own WTP. **Quantitatively, for each 1 JPY that a participant overestimates peers' WTP_1, their own WTP_1 increases by about 0.6–0.7 JPY; for each 1 JPY overestimation of peers' WTP_2, their own WTP_2 increases by about 0.4–0.5 JPY.** This pattern is consistent with a

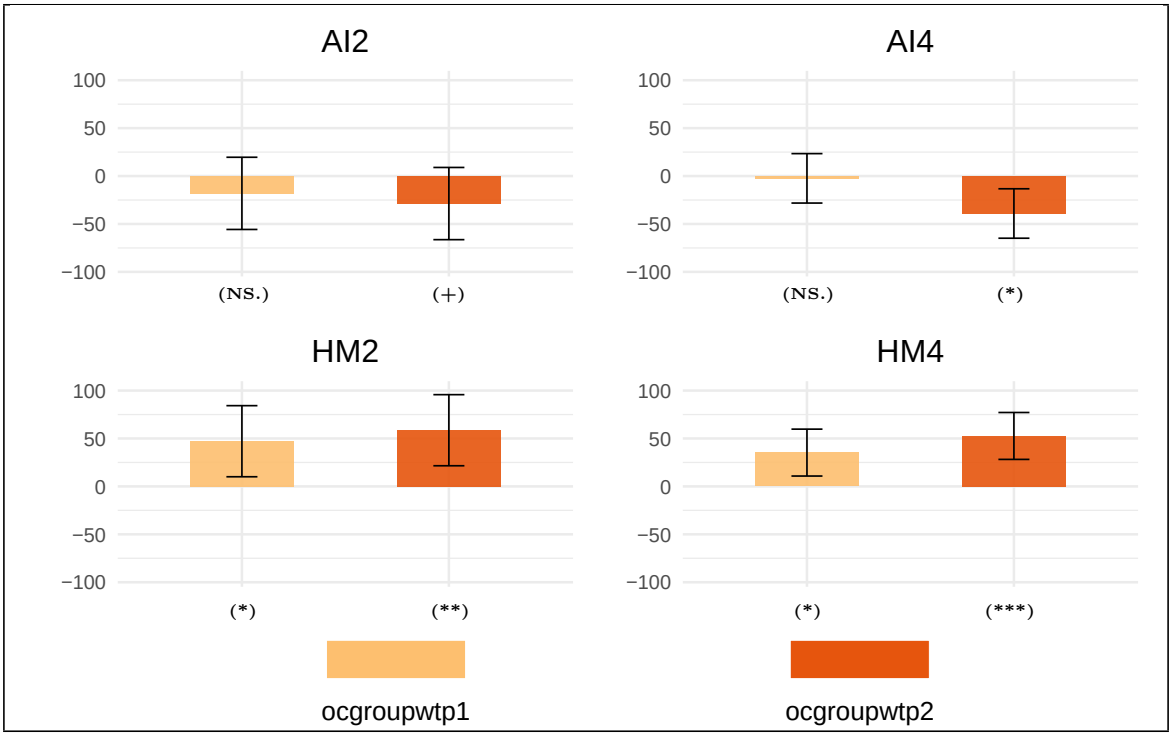


Figure 17: Overestimation of Peers' WTP

Note: The symbols shown in parentheses below each bar report the one-sample Wilcoxon test against zero for that variable. The symbols $^{+}$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. Error bars denote 95% confidence intervals across participants.

projection/false–consensus effect (Ross et al., 1977; Marks and Miller, 1987), which means that participants who believe others will pay more (less) also tend to report higher (lower) WTP themselves, suggesting that social beliefs may amplify or dampen adoption of assistance. Interactions with treatments are generally insignificant.

5.3 How do people detect deepfake news?

Detection Strategies. In the posterior-belief survey, we asked a multiple-choice question with multiple selections allowed about how people detect deepfakes (see Online Appendix D.4, Q6). Participants could select any of the following: (i) detecting factual errors; (ii) noticing unnatural grammar or wording; (iii) noticing inconsistent or disjointed context or logic.

Among 158 participants, 27.2% selected option (i), 88.6% selected option (ii), and 82.9% selected option (iii), indicating that participants relied more on language cues and internal consistency than on explicit fact-checking. In addition to the predefined cues (factual errors, grammatical anomalies, and logical inconsistencies), open-ended responses most often mentioned (i) external verification (e.g., checking named entities and timelines) and (ii) stylistic regularities suggestive of formulaic writing (e.g., generic conclusions and uniform sentence structures). Participants also reported monitoring the internal consistency of details (numbers, places, institutional names) and attending to "AI-like" meta-features (overly neutral tone and few concrete examples).

Determinants of Detection Accuracy. We then estimated OLS regressions of detection accuracy on demographics, the actual AI proportion of the news item, and the score of the matrix quiz (`matrixquiz`; integer 0–7). Results are reported in Tables A.9–A.10 of Online Appendix A. The main findings are as follows: (i) the negative and statistically significant coefficient on `female` indicates a gender gap in detection accuracy, with female participants exhibiting lower average accuracy; (ii) a **higher AI proportion** is associated with lower detection accuracy and smaller within-task improvement, and this association is stronger for items expanded by GPT-4o, suggesting that more sophisticated and pervasive AI-generated content reduces both the detectability of deepfakes and participants' ability to adapt through experience; and (iii) the `matrixquiz` score is not statistically significant, indicating little association between reasoning ability and deepfake detection in this setting.

ℂℍ𝔸𝕋 **Logs.** To examine how participants used ℂℍ𝔸𝕋 for detection, we first translated the chat logs into English using GPT-4o via API.[18] We then applied basic natural language processing (NLP) and constructed word clouds. Figure 18 show the word

[18]The Python translation script, the original Japanese chat logs, and the English translations are available at https://github.com/kazewindser/GithubAppendixPATDA.

cloud from conversations with the AI detector (with AI replies and pasted article text removed) in Panel (a), that from Peer chat (with pasted article text removed) in Panel (b). As a robustness check, we also processed the original Japanese logs and generated Japanese word clouds; the corresponding results are reported in Online Appendix A.3.

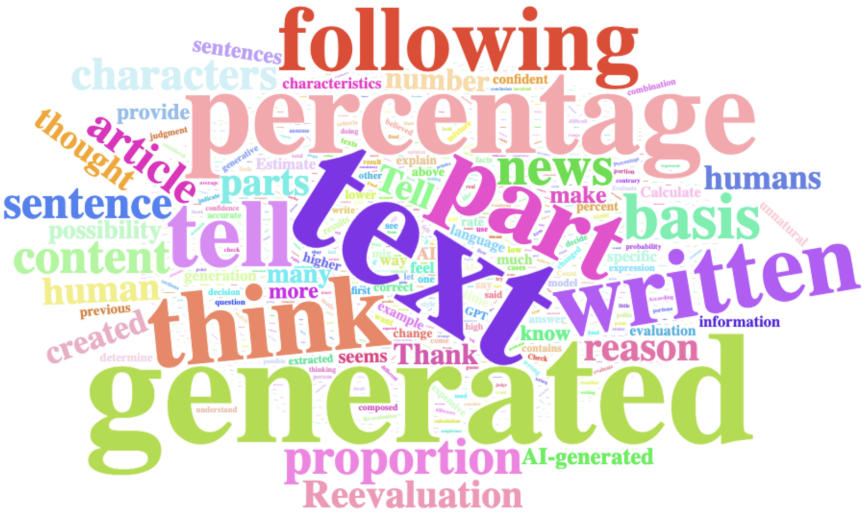
In the AI-detector condition, the 20 most frequent words were: *text, generated, percentage, think, part, written, tell, following, basis, news, proportion, content, article, sentence, characters, reason, parts, reevaluation, human, tell.*

By contrast, in the Peer chat condition, the top words were: *think, thought, many, understand, percentage, seems, set, first, see, half, human, same, much, sentence, make, changed, feel, unnatural, part, time.*

These patterns suggest different linguistic emphases. With AI detector, participants more often use *task/evidence* terms (e.g., *generated, percentage, proportion, content, article, characters*), whereas with peers they use more *introspective/metacognitive* language (e.g., *think, thought, understand, seems, see, feel, unnatural*). Taken together, this points to distinct collaboration modes: under AI collaboration, participants treat the assistant as a **"rule retriever"** — prompting evidence-seeking and structured queries—whereas under peer collaboration, they use a **"social calibrator"**, sharing uncertainty and monitoring their own reasoning.

6 Conclusion

Using a 2×2 between-subjects design, we ran a lab experiment to compare participants' WTP for access to a ChatGPT-based AI detector with their WTP to collaborate with human peers on a deepfake detection task. We varied task difficulty by including deepfake news generated by two different LLMs (GPT-4o vs. GPT-2). We study how the treatment (AI detector vs. Peer chat), task difficulty, and performance feedback shape WTP. Our aim is to identify the relative, private value of human–AI collaboration



(a) The AI detector condition

(b) The Peer chat condition

Figure 18: Word Cloud of $\mathbb{CHAT}$ in (a) the AI detector condition and (b) the Peer chat condition

in the context of using AI tools to manage risks created by AI itself.

Participants performed worse when detecting GPT-4o news than GPT-2 news. They were also more willing to pay for access to the AI detector than to chat with human peers, even though access to the detector did not significantly improve accuracy. We also find a feedback effect: after doing the task and seeing feedback, participants increased their WTP, with the change concentrated in WTP for the AI detector and driven by positive feedback. By contrast, task difficulty shows little impact on WTP: although GPT-4o news is harder to detect than GPT-2 news for both AI and humans, this difficulty gap does not meaningfully change WTP or its post-feedback shift. Participants overpaid on average for both AI and Human assistence, but more so for the AI. Ex-ante preferences did not predict WTP, but after experiencing the task, relative demand shifted, consistent with learning rather than fixed tastes. Beliefs were miscalibrated: participants underestimated their own and peers' performance and tended to rate AI relatively higher as tasks became harder.

This study has several limitations. First, while the lab setting provides tight control (e.g., preventing the use of outside AI tools), it limits external validity; each participant evaluated only 22 articles over a short horizon. Second, our "detector" is a ChatGPT-based implementation rather than a certified stand-alone tool, so brand and interface preferences may affect WTP beyond pure accuracy. Third, we benchmarked deepfakes from only two LLMs and used a single detector design, focusing on text rather than images or video; robustness may therefore be limited and results sensitive to model drift. These limits motivate follow-ups with field deployments, brand-blinded interfaces, multiple detector back ends, and a richer matrix of generators and content types.

Within these boundaries, several policy-relevant lessons remain. First, both humans and current AI detectors are limited at deepfake detection. Policymakers and institutions should treat deepfakes as a rising harm: raise awareness, teach practical red flags, and caution against over-reliance on detectors. Second, regulation should

focus on quality, transparency, and user protection: require clear labels and truthful accuracy ranges, set minimum performance floors for high-stakes uses, restrict inflated marketing claims, and move toward outcome-based pricing and independent audits. Third, decisions should not be fully delegated to AI. For ambiguous or high-risk cases, keep humans in the loop—pair detectors with peer review or expert checks to reduce bias and drift. Taken together, we hope this study takes a small step toward balancing rapid AI progress with safeguards against the risks it creates.

References

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

Alexander, S. (2025): "Deepfake Cyberbullying: The Psychological Toll on Students and Institutional Challenges of AI-Driven Harassment," *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 98, 36–50.

Amin, A., Y. Hong, and B. Mazhar (2025): "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, 321–339.

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

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

Becker, G. M., M. H. DeGroot, and J. Marschak (1964): "Measuring utility by a single-response sequential method," *Behavioral science*, 9, 226–232.

Ben David, D., Y. S. Resheff, and T. Tron (2021): "Explainable AI and adoption

of financial algorithmic advisors: an experimental study," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 390–400.

Bergin, J. (2025a): "Over a dozen unis are using AI to catch AI — and getting it wrong," Accessed: 2025-11-01.

——— (2025b): "University wrongly accuses students of using artificial intelligence to cheat," Accessed: 2025-11-01.

Bittle, K. and O. El-Gayar (2025): "Generative AI and academic integrity in higher education: A systematic review and research agenda," *Information*, 16, 296.

Brodeur, A., D. Valenta, A. Marcoci, J. P. Aparicio, D. Mikola, B. Barbarioli, R. Alexander, L. Deer, T. Stafford, L. Vilhuber, et al. (2025): "Comparing Human-Only, AI-Assisted, and AI-Led Teams on Assessing Research Reproducibility in Quantitative Social Science," Tech. rep., I4R Discussion Paper Series.

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

Chacon, A., E. E. Kausel, T. Reyes, and S. Trautmann (2025): "Preventing algorithm aversion: People are willing to use algorithms with a learning label," *Journal of Business Research*, 187, 115032.

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

Chaka, C. (2024): "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, 115–126.

Chakravarti, A. (2023): "Oh god! Open AI tool that identifies text written by ChatGPT believes Bible was written by AI," Updated: 2023-02-02.

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

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

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

Ching, D., J. Twomey, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy (2025): "Can deepfakes manipulate us? Assessing the evidence via a critical scoping review," *PLoS One*, 20, e0320124.

Chong, A. T. Y., H. N. Chua, M. B. Jasser, and R. T. Wong (2023): "Bot or human? detection of deepfake text with semantic, emoji, sentiment and linguistic features," in *2023 IEEE 13th International Conference on System Engineering and Technology (ICSET)*, IEEE, 205–210.

Choudhary, V., A. Marchetti, Y. R. Shrestha, and P. Puranam (2025): "Human-AI ensembles: When can they work?" *Journal of Management*, 51, 536–569.

Condon, D. M. and W. Revelle (2014): "The international cognitive ability resource: Development and initial validation of a public-domain measure," *Intelligence*, 43, 52–64.

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

Donahue, K., A. Chouldechova, and K. Kenthapadi (2022): "Human-algorithm collaboration: Achieving complementarity and avoiding unfairness," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639–1656.

Doshi, A. R. and O. P. Hauser (2024): "Generative AI enhances individual creativity but reduces the collective diversity of novel content," *Science advances*, 10, eadn5290.

Dratsch, T., X. Chen, M. Rezazade Mehrizi, R. Kloeckner, A. Mähringer-Kunz, M. Püsken, B. Baessler, S. Sauer, D. Maintz, and D. Pinto dos Santos (2023): "Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance," *Radiology*, 307, e222176.

Duckers, S., J. Schepers, T. Zwienenberg, P. Verheijen, and M. Westenenk (2024): "AI-Human Team Compositions in Advisory Services: Customer Attitudes, Attribution Effects, and the Mitigating Role of Discounts," .

Ebbers, F., J. Zibuschka, C. Zimmermann, and O. Hinz (2021): "User preferences for privacy features in digital assistants," *Electronic Markets*, 31, 411–426.

Edwards, B. (2023): "Why AI writing detectors don't work," Explains false positives such as the U.S. Constitution being flagged as AI-generated.

Gaebler, J. D., S. Goel, A. Huq, and P. Tambe (2024): "Auditing large language

models for race & gender disparities: Implications for artificial intelligence-based hiring," *Behavioral Science & Policy*, 10, 46–55.

Garde, A., S. Suratkar, and F. Kazi (2022): "AI based deepfake detection," in *2022 IEEE 1st International Conference on Data, Decision and Systems (ICDDS)*, IEEE, 1–6.

Gino, F. and D. A. Moore (2007): "Effects of task difficulty on use of advice," *Journal of Behavioral Decision Making*, 20, 21–35.

Goh, E., R. Gallo, J. Hom, E. Strong, Y. Weng, H. Kerman, J. A. Cool, Z. Kanjee, A. S. Parsons, N. Ahuja, et al. (2024): "Large language model influence on diagnostic reasoning: a randomized clinical trial," *JAMA network open*, 7, e2440969–e2440969.

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

Greshake, K., S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz (2023): "Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection," in *Proceedings of the 16th ACM workshop on artificial intelligence and security*, 79–90.

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

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

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

Gupta, S., U. Sharma, and L. Vardari (2025): "The Influence of Deepfake Technology on Political Affairs," in *Mastering Deepfake Technology: Strategies for Ethical Management and Security*, River Publishers, 147–162.

Hallikaar, V. (2025): "Western N.Y. student's AI use accusation questions validity, raises concerns," Accessed: 2025-11-01.

Harrison, G. W. and E. E. Rutström (2008): "Experimental evidence on the existence of hypothetical bias in value elicitation methods," *Handbook of experimental economics results*, 1, 752–767.

Hemmer, P., M. Schemmer, N. Kühl, M. Vössing, and G. Satzger (2025): "Complementarity in human-AI collaboration: Concept, sources, and evidence," *European Journal of Information Systems*, 1–24.

Ioku, T., J. Song, and E. Watamura (2024): "Trade-offs in AI assistant choice: Do consumers prioritize transparency and sustainability over AI assistant performance?" *Big Data & Society*, 11, 20539517241290217.

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

Islam, R. and O. M. Moushi (2024): "Gpt-4o: The cutting-edge advancement in multimodal llm," *Authorea Preprints*.

Jo, H. (2025): "Uncovering the reasons behind willingness to pay for ChatGPT-4 premium," *International Journal of Human–Computer Interaction*, 41, 994–1009.

Kar, S. K., T. Bansal, S. Modi, and A. Singh (2024): "How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools," *Indian Journal of Psychological Medicine*, 02537176241247934.

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

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018): "Human decisions and machine predictions," *The quarterly journal of economics*, 133, 237–293.

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

Knilans, G. (2024): "The Dark Side of AI Detectors: Why Accuracy Is Not Guaranteed," https://www.tradepressservices.com/ai-detectors/, trade Press Services blog post.

Köbis, N., Z. Rahwan, R. Rilla, B. I. Supriyatno, C. Bersch, T. Ajaj, J.-F. Bonnefon, and I. Rahwan (2025): "Delegation to artificial intelligence can increase dishonest behaviour," *Nature*, 1–9.

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

König, P. D., S. Wurster, and M. B. Siewert (2022): "Consumers are willing to pay a price for explainable, but not for green AI. Evidence from a choice-based conjoint analysis," *Big Data & Society*, 9, 20539517211069632.

Kuberska, D. and K. Klaudia (2025): "To Pay or Not to Pay? Investigating Students' Willingness to Pay for ChatGPT," *Olsztyn Economic Journal*, 20, 169–183.

Küper, A. and N. Krämer (2025): "Psychological traits and appropriate reliance: Factors shaping trust in AI," *International Journal of Human–Computer Interaction*, 41, 4115–4131.

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

Lee, D. (2024): "Q&A: The Increasing Difficulty of Detecting AI- Versus Human-Generated Text," *Penn State News*, accessed 18 May 2025.

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

Lee, M. H. and C. J. Chew (2023): "Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making," *Proceedings of the ACM on Human-Computer Interaction*, 7, 1–22.

Li, S. (2025): "The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation," *Digital Society & Virtual Governance*, 1, 72–88.

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

Lu, Z., D. Wang, and M. Yin (2024): "Does more advice help? the effects of second opinions in AI-assisted decision making," *Proceedings of the ACM on Human-Computer Interaction*, 8, 1–31.

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

Lupa-Wójcik, I. (2024): "Students' willingness to pay for access to ChatGPT," .

Marks, G. and N. Miller (1987): "Ten years of research on the false-consensus effect: An empirical and theoretical review." *Psychological bulletin*, 102, 72.

Masood, M., M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik (2023): "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied intelligence*, 53, 3974–4026.

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

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

Passi, S. and M. Vorvoreanu (2022): "Overreliance on AI literature review," *Microsoft Research*, 339, 340.

Ploug, T., A. Sundby, T. B. Moeslund, and S. Holm (2021): "Population preferences for performance and explainability of artificial intelligence in health care: choice-based conjoint survey," *Journal of Medical Internet Research*, 23, e26611.

Ramluckan, T. (2024): "Deepfakes: The legal implications," in *International Conference on Cyber Warfare and Security*, Academic Conferences International Limited, vol. 19, 282–288.

Ross, L., D. Greene, and P. House (1977): "The "false consensus effect": An egocentric bias in social perception and attribution processes," *Journal of experimental social psychology*, 13, 279–301.

Saharan, S., M. Sharma, M. Khiria, and A. Gupta (2025): "Human-AI Collaboration in Moderating AI-Generated Harmful Content," in *Content Moderation in the Age of AI*, IGI Global Scientific Publishing, 123–166.

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

Sareen, M. (2022): "Threats and challenges by DeepFake technology," in *DeepFakes*, CRC Press, 99–113.

Sawada, K., T. Zhao, M. Shing, K. Mitsui, A. Kaga, Y. Hono, T. Wakatsuki, and K. Mitsuda (2024): "Release of Pre-Trained Models for the Japanese Language," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13898–13905, https://arxiv.org/abs/2404.01657.

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

Somoray, K., D. J. Miller, and M. Holmes (2025): "Human Performance in Deepfake Detection: A Systematic Review," *Human Behavior and Emerging Technologies*, 2025, 1833228.

Sophia, L. (2025): "The Social Harms of AI-Generated Fake News: Addressing Deepfake and AI Political Manipulation," *Digital Society & Virtual Governance*, 1, 72–88.

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

The Advertiser (2025): "'Robocheating' fiasco saw ACU students falsely accused of using AI by an unreliable tool," Accessed: 2025-11-01.

The Courier-Mail (2024): "'Impossible': New AI detection tool slammed by experts," Accessed: 2025-11-01.

Uchendu, A., J. Lee, H. Shen, T. Le, D. Lee, et al. (2023): "Does human collaboration enhance the accuracy of identifying llm-generated deepfake texts?" in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 11, 163–174.

Umbach, R., N. Henry, G. F. Beard, and C. M. Berryessa (2024): "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*, 1–20.

Vaccaro, M., A. Almaatouq, and T. Malone (2024): "When combinations of humans and AI are useful: A systematic review and meta-analysis," *Nature Human Behaviour*, 8, 2293–2303.

Varri, D. B. S. V. et al. (2025): "Human-AI collaboration in healthcare security," .

Vasconcelos, H., M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna (2023): "Explanations can reduce overreliance on ai systems during decision-making," *Proceedings of the ACM on Human-Computer Interaction*, 7, 1–38.

von Wedel, P. and C. Hagist (2022): "Physicians' preferences and willingness to pay for artificial intelligence-based assistance tools: a discrete choice experiment among german radiologists," *BMC Health Services Research*, 22, 398.

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

Whittaker, L., R. Mulcahy, R. Russell-Bennett, K. Letheren, and J. Kietzmann (2025): "Examining Consumer Appraisals of Deepfake Advertising and Disclosure: Show Deepfakes as "Real Life" or Say They're "Just Fantasy"?" *Journal of Advertising Research*, 1–22.

Wilson, K. and A. Caliskan (2024): "Gender, race, and intersectional bias in resume screening via language model retrieval," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 1578–1590.

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

Zellers, R., A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi (2019): "Defending against neural fake news," *Advances in neural information processing systems*, 32.

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

Zhang, Q., C. Gao, D. Chen, Y. Huang, Y. Huang, Z. Sun, S. Zhang, W. Li, Z. Fu, Y. Wan, et al. (2024a): "LLM-as-a-coauthor: Can mixed human-written and machine-generated text be detected?" *arXiv preprint arXiv:2401.05952*.

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

Zhang, Z., B. Guo, and T. Li (2024c): "Privacy Leakage Overshadowed by Views of AI: A Study on Human Oversight of Privacy in Language Model Agent," *arXiv preprint arXiv:2411.01344*.

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

Zhao, T. and K. Sawada (2021): "rinna/japanese-gpt2-medium," .

Zhu, F. and W. Zou (2023): "The role of generative AI in human creative processes: Experimental evidence," *Available at SSRN 4676053*.

Ziegler, A., E. Kalliamvakou, X. A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam, and E. Aftandilian (2024): "Measuring github copilot's impact on productivity," *Communications of the ACM*, 67, 54–63.

Online Appendix to "Paying AI to Detect AI"

Yuhao Fu* Nobuyuki Hanaki†

November 10, 2025

Contents

*Graduate School of Economics, the University of Osaka. E-mail: kazewindser@gmail.com
†Corresponding author. Institute of Social and Economic Research, the University of Osaka, and University of Limassol. E-mail: nobuyuki.hanaki@iser.osaka-u.ac.jp

A Additional Analyses

A.1 Regressions

Table A.1: OLS Regressions of WTP on Demographics

Dep. Var.	WTP1		WTP2		RltWTPup	
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	115.984***	115.984***	154.017***	154.017***	0.110	0.110
	(17.634)	(17.634)	(19.908)	(19.908)	(0.082)	(0.082)
gpt4news	15.940	15.940	16.374	16.374	0.015	0.015
	(18.264)	(18.264)	(20.620)	(20.620)	(0.085)	(0.085)
age	−4.536	−4.536	−0.867	−0.867	0.0004	0.0004
	(5.007)	(5.007)	(5.653)	(5.653)	(0.023)	(0.023)
female	35.595^{+}	35.595^{+}	13.186	13.186	−0.091	−0.091
	(20.129)	(20.129)	(22.724)	(22.724)	(0.094)	(0.094)
edulevel	39.957	39.957	14.915	14.915	0.006	0.006
	(26.747)	(26.747)	(30.196)	(30.196)	(0.124)	(0.124)
lanjp	−42.101	−42.101	−38.060	−38.060	−0.163	−0.163
	(44.957)	(44.957)	(50.754)	(50.754)	(0.216)	(0.216)
engr	−30.064	−30.064	−58.251*	−58.251*	−0.174	−0.174
	(23.796)	(23.796)	(26.864)	(26.864)	(0.110)	(0.110)
linguis	−28.759	−28.759	8.159	8.159	0.152	0.152
	(25.315)	(25.315)	(28.579)	(28.579)	(0.117)	(0.117)
freqGPT	2.246	2.246	1.491	1.491	−0.009	−0.009
	(4.235)	(4.235)	(4.781)	(4.781)	(0.020)	(0.020)
exprog	5.814	5.814	−0.201	−0.201	−0.048	−0.048
	(21.458)	(21.458)	(24.225)	(24.225)	(0.100)	(0.100)
GPTplus	−47.892	−47.892	−79.147^{+}	−79.147^{+}	−0.034	−0.034
	(36.095)	(36.095)	(40.749)	(40.749)	(0.168)	(0.168)
Constant	207.315	207.315	133.274	133.274	0.165	0.165
	(133.571)	(133.571)	(150.794)	(150.794)	(0.630)	(0.630)
Observations	158	158	158	158	156	156
R^2	0.259	0.259	0.323	0.323	0.064	0.064

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. `lanjp`, `engr`, and `linguis` are indicator variables: `lanjp` = 1 if the participant is a Japanese native speaker; `engr` = 1 if the participant's major is engineering; `linguis` = 1 if the participant's major is linguistics or humanities. Numbers in parentheses represent standard errors.

Table A.2: OLS Regressions of WTP on Beliefs

Dep. Var.	WTP1		WTP2		RltWTPup	
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	107.415***	104.112***	147.841***	148.918***	0.106	0.119
	(18.766)	(17.476)	(21.036)	(19.048)	(0.086)	(0.079)
gpt4news	16.685	16.034	17.900	21.853	0.041	0.053
	(17.814)	(17.606)	(19.949)	(19.250)	(0.081)	(0.080)
pcvdDiff	−5.567	−4.657	3.060	4.846	0.036	0.038
	(13.122)	(13.043)	(14.957)	(14.454)	(0.061)	(0.060)
pcvdFami			−15.023	−11.784	−0.015	−0.009
			(9.089)	(8.823)	(0.037)	(0.037)
pcvdDanger	−9.937	−9.864	20.973	21.163	0.124*	0.125*
	(12.747)	(12.734)	(14.384)	(13.991)	(0.058)	(0.058)
pcvdRisk	−9.639	−8.258	−15.686	−10.971	−0.010	0.001
	(9.753)	(9.849)	(10.910)	(10.735)	(0.045)	(0.045)
prefCHATprior	−5.306		1.562		0.022	
	(10.786)		(12.207)		(0.050)	
prefCHATpost		8.011		34.247**		0.078
		(10.784)		(11.861)		(0.049)
Constant	191.880**	172.028*	60.941	−4.676	−0.706*	−0.834*
	(69.059)	(70.388)	(81.370)	(80.845)	(0.332)	(0.336)
Observations	158	158	158	158	156	156
R^2	0.211	0.212	0.296	0.333	0.055	0.069

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. `pcvdDiff`, `pcvdFami`, `pcvdDanger`, and `pcvdRisk` are coded based on participants' responses in the posterior-belief survey (Online Appendix D.4, Questions 2–4). They measure, respectively, (i) perceived task difficulty, (ii) perceived familiarity with the news content, (iii) perceived social danger of deepfake news, and (iv) perceived risk associated with AI tools. Each variable is coded on a five-point scale (1–5), where larger values indicate higher perceived difficulty, greater familiarity, stronger perceived social danger, and higher perceived risk of AI tools. Numbers in parentheses represent standard errors.

Table A.3: First Self-Overconfidence & WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	102.670***	108.315***	102.033***	142.543***	153.507***	145.130***
	(17.290)	(25.889)	(17.391)	(19.608)	(29.344)	(19.563)
gpt4news	14.749	15.069	6.159	24.348	24.969	59.237*
	(17.244)	(17.330)	(25.556)	(19.556)	(19.643)	(28.748)
ocslfaccu1	−83.738+	−96.527	−63.100	−62.689	−87.529	−146.509+
	(48.735)	(65.469)	(66.572)	(55.269)	(74.205)	(74.886)
inAI × ocslfaccu1		28.947			56.223	
		(98.575)			(111.729)	
gpt4news × ocslfaccu1			−44.606			181.169
			(97.728)			(109.934)
Constant	76.639***	74.238***	81.090***	60.499**	55.835*	42.420+
	(17.617)	(19.470)	(20.176)	(19.979)	(22.068)	(22.696)
Observations	158	158	158	158	158	158
R^2	0.211	0.212	0.212	0.272	0.274	0.285

Note: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.4: Second Self-Overconfidence & WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	106.802***	118.786***	106.744***	145.390***	159.672***	145.977***
	(17.317)	(18.149)	(17.371)	(19.597)	(20.509)	(19.346)
gpt4news	17.081	17.481	15.449	25.274	25.751	41.639*
	(17.606)	(17.435)	(18.830)	(19.924)	(19.702)	(20.970)
ocslfaccu2	−53.537	−145.173*	−40.881	−27.624	−136.835+	−154.529*
	(47.727)	(65.586)	(69.717)	(54.011)	(74.114)	(77.640)
inAI × ocslfaccu2		187.486*			223.444*	
		(93.043)			(105.141)	
gpt4news × ocslfaccu2			−23.917			239.812*
			(95.772)			(106.656)
Constant	86.040***	79.079***	87.323***	68.894***	60.597**	56.033**
	(16.165)	(16.376)	(17.009)	(18.294)	(18.505)	(18.942)
Observations	158	158	158	158	158	158
R^2	0.203	0.223	0.203	0.268	0.289	0.291

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.5: First Peer Overconfidence & WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	105.089***	137.258***	105.043***	143.467***	162.234***	143.585***
	(17.420)	(25.908)	(17.478)	(19.622)	(29.376)	(19.679)
gpt4news	14.297	14.145	11.024	25.006	24.917	33.444
	(17.472)	(17.371)	(26.165)	(19.680)	(19.697)	(29.461)
ocgroupaccu1	−23.755	−113.203	−15.540	−50.047	−102.231	−71.222
	(56.247)	(77.444)	(74.571)	(63.356)	(87.812)	(83.964)
inAI × ocgroupaccu1		185.947+			108.482	
		(111.371)			(126.282)	
gpt4news × ocgroupaccu1			−19.131			49.311
			(113.533)			(127.833)
Constant	87.711***	73.611***	89.286***	63.137**	54.911*	59.077*
	(18.248)	(20.012)	(20.554)	(20.555)	(22.692)	(23.143)
Observations	158	158	158	158	158	158
R^2	0.197	0.212	0.197	0.269	0.273	0.270

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.6: Second Peer Overconfidence & WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	105.706***	116.830***	105.345***	144.811***	154.863***	143.604***
	(17.403)	(18.316)	(17.454)	(19.631)	(20.742)	(19.536)
gpt4news	13.438	13.289	16.991	23.350	23.216	35.215
	(17.894)	(17.760)	(19.055)	(20.185)	(20.112)	(21.328)
ocgroupaccu2	1.813	−117.382	−31.774	1.582	−106.131	−110.578
	(60.736)	(88.885)	(86.084)	(68.512)	(100.655)	(96.351)
inAI × ocgroupaccu2		215.157+			194.431	
		(117.908)			(133.521)	
gpt4news × ocgroupaccu2			67.071			223.975
			(121.547)			(136.045)
Constant	92.049***	84.963***	89.294***	72.057***	65.654***	62.857**
	(16.465)	(16.797)	(17.241)	(18.573)	(19.021)	(19.297)
Observations	158	158	158	158	158	158
R^2	0.196	0.213	0.198	0.266	0.276	0.279

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.7: AI-Overconfidence & WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	102.814***	101.301***	102.193***	144.583***	144.879***	141.861***
	(17.471)	(17.723)	(17.868)	(19.792)	(20.097)	(20.213)
gpt4news	1.772	0.452	1.065	22.415	22.672	19.318
	(20.094)	(20.281)	(20.543)	(22.764)	(22.998)	(23.239)
ocAI	53.508	31.049	46.261	4.741	9.125	−27.004
	(46.235)	(61.619)	(61.626)	(52.379)	(69.876)	(69.714)
inAI × ocAI		44.609			−8.707	
		(80.675)			(91.485)	
gpt4news × ocAI			16.812			73.649
			(94.142)			(106.497)
Constant	97.798***	98.681***	97.505***	72.428***	72.256***	71.142***
	(16.136)	(16.251)	(16.271)	(18.281)	(18.429)	(18.406)
Observations	158	158	158	158	158	158
R^2	0.203	0.205	0.203	0.266	0.266	0.269

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.8: Overestimation of Peers' WTP and Own WTP

Dep. Var.	WTP1			WTP2		
	(1)	(2)	(3)	(4)	(5)	(6)
inAI	141.096***	139.690***	141.580***	184.238***	186.554***	182.402***
	(13.752)	(13.944)	(13.625)	(20.102)	(20.192)	(20.317)
gpt4news	12.349	11.741	8.062	27.044	26.837	28.269
	(13.349)	(13.406)	(13.398)	(18.349)	(18.335)	(18.469)
ocgroupwtp1	0.699***	0.654***	0.603***			
	(0.067)	(0.096)	(0.083)			
inAI × ocgroupwtp1		0.089				
		(0.135)				
gpt4news × ocgroupwtp1			0.267*			
			(0.135)			
ocgroupwtp2				0.439***	0.544***	0.483***
				(0.093)	(0.132)	(0.113)
inAI × ocgroupwtp2					−0.208	
					(0.185)	
gpt4news × ocgroupwtp2						−0.119
						(0.174)
Constant	64.021***	66.177***	65.181***	45.627**	39.916*	45.887**
	(12.096)	(12.558)	(11.996)	(17.123)	(17.852)	(17.157)
Observations	158	158	158	158	158	158
R^2	0.527	0.528	0.539	0.360	0.365	0.362

Note: $^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Numbers in parentheses represent standard errors.

Table A.9: Determinants of Detection Accuracy: Demographics & AIpro

Dep. Var.	accu1				accu2			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
inAI	−0.009	−0.009	0.033^{+}	−0.013	−0.013	−0.013	0.063**	−0.015
	(0.008)	(0.008)	(0.020)	(0.009)	(0.009)	(0.009)	(0.021)	(0.009)
gpt4news	−0.043***	0.034^{+}	−0.043***	−0.048***	−0.076***	0.041*	−0.076***	−0.083***
	(0.008)	(0.019)	(0.008)	(0.009)	(0.010)	(0.020)	(0.010)	(0.011)
female	−0.019*	−0.019*	−0.019*		−0.018^{+}	−0.018^{+}	−0.018^{+}	
	(0.009)	(0.009)	(0.009)		(0.010)	(0.010)	(0.010)	
matrixquiz				0.001				0.001
				(0.002)				(0.003)
AIpro	−0.002***	−0.001***	−0.002***	−0.002***	−0.001***	0.0001	−0.0003	−0.001***
	(0.0002)	(0.0003)	(0.0003)	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0002)
age	0.001	0.001	0.001	0.0005	0.004^{+}	0.004^{+}	0.004^{+}	0.003
	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)
lanjp	0.033^{+}	0.033^{+}	0.033^{+}	0.010	0.057^{+}	0.057^{+}	0.057^{+}	0.004
	(0.020)	(0.020)	(0.020)	(0.022)	(0.031)	(0.031)	(0.031)	(0.032)
engr	0.002	0.002	0.002	0.004	−0.003	−0.003	−0.003	−0.008
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)	(0.010)	(0.011)
linguis	0.009	0.009	0.009	0.004	−0.008	−0.008	−0.008	−0.011
	(0.011)	(0.011)	(0.011)	(0.012)	(0.015)	(0.015)	(0.015)	(0.015)
edulevel	0.001	0.001	0.001	0.001	−0.013	−0.013	−0.013	−0.013
	(0.011)	(0.011)	(0.011)	(0.012)	(0.013)	(0.013)	(0.013)	(0.013)
gpt4news × AIpro		−0.002***				−0.002***		
		(0.0004)				(0.0004)		
inAI × AIpro			−0.001*				−0.002***	
			(0.0004)				(0.0004)	
Constant	0.790***	0.749***	0.769***	0.821***	0.689***	0.627***	0.651***	0.784***
	(0.069)	(0.069)	(0.069)	(0.068)	(0.078)	(0.077)	(0.076)	(0.080)
Obs.	3476	3476	3476	3476	3476	3476	3476	3476
R^2	0.093	0.104	0.096	0.097	0.053	0.078	0.064	0.060
Clusters	158	158	158	158	158	158	158	158

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

Table A.10: Determinants of Performance Improvment: Demographics & AIpro

Dep. Var.	PRE				accuUP			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
inAI	−0.134	−0.134	0.062	−0.086	−0.045	−0.045	0.047	−0.049
	(0.099)	(0.100)	(0.167)	(0.107)	(0.053)	(0.053)	(0.087)	(0.059)
gpt4news	−0.005	0.201	−0.006	0.035	−0.260***	−0.182*	−0.261***	−0.284***
	(0.104)	(0.171)	(0.104)	(0.120)	(0.057)	(0.089)	(0.057)	(0.058)
female	0.007	0.007	0.006		0.053	0.053	0.053	
	(0.108)	(0.108)	(0.108)		(0.062)	(0.062)	(0.062)	
matrixquiz				0.005				−0.001
				(0.031)				(0.016)
AIpro	0.007***	0.009***	0.009***	0.006***	0.005***	0.006***	0.006***	0.004***
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
age	0.023	0.023	0.023	0.018	0.007	0.007	0.007	−0.004
	(0.026)	(0.026)	(0.026)	(0.027)	(0.016)	(0.016)	(0.016)	(0.015)
lanjp	0.039	0.040	0.041	−0.003	0.211	0.211	0.211	−0.009
	(0.150)	(0.150)	(0.150)	(0.195)	(0.164)	(0.164)	(0.164)	(0.136)
engr	−0.062	−0.063	−0.063	−0.089	−0.067	−0.067	−0.067	−0.141*
	(0.128)	(0.129)	(0.128)	(0.126)	(0.074)	(0.074)	(0.074)	(0.069)
linguis	−0.107	−0.108	−0.108	−0.096	−0.157*	−0.157*	−0.157*	−0.199*
	(0.152)	(0.152)	(0.152)	(0.168)	(0.077)	(0.077)	(0.077)	(0.081)
edulevel	−0.220	−0.219	−0.219	−0.222	−0.040	−0.040	−0.040	−0.038
	(0.144)	(0.144)	(0.144)	(0.155)	(0.080)	(0.080)	(0.080)	(0.083)
gpt4news × AIpro		−0.004+				−0.002		
		(0.002)				(0.002)		
inAI × AIpro			−0.004				−0.002	
			(0.002)				(0.002)	
Constant	−1.234+	−1.341*	−1.333*	−1.110	−0.473	−0.515	−0.519	0.069
	(0.654)	(0.669)	(0.658)	(0.719)	(0.471)	(0.478)	(0.474)	(0.415)
Obs.	3476	3476	3476	3476	3476	3476		
R^2	0.093	0.104	0.096	0.097	0.053	0.078	0.064	0.060
Clusters	158	158	158	158	158	158		

Note: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Models (1)–(4) are OLS regressions; Models (5)–(8) are Probit regressions. Standard errors have been corrected for within-subjects clustering effects to account for the non-independence of observations from the same participant. Numbers in parentheses represent standard errors.

A.2 Potential Profit Based on the Expected Payment

As a robustness check to **Section 5.1**, we use the *expected payment under the BDM draw* as the effective price, $ExpWTP_{rn2} = \frac{1+WTP_{rn2}}{2}$. For participant i, redefine the *potential profit matrix* $\mathbf{\Phi}_i^* \in \mathbb{R}^{22\times 22}$, whose entry $\phi^*_{i,rn1,rn2}$ is the potential profit if the first draw is round $rn1$ and the second draw is round $rn2$ ($rn1, rn2 \in \{1, \ldots, 22\}$):

$$\begin{aligned}\phi^*_{i,rn1,rn2} \;=\; & 0.2 \cdot \max\Big\{0,\, 2300 - 0.3 \cdot (AIpro^*_{rn1} - 1stResp_{i,rn1})^2\Big\} \\ & + 0.8 \cdot \max\Big\{0,\, 2300 - 0.3 \cdot (AIpro^*_{rn2} - 2ndResp_{i,rn2})^2\Big\} \\ & - inChat_{i,rn2} \cdot ExpWTP_{i,rn2}\end{aligned}$$

Then, figure A.1 presents the heatmap of average potential profits across the four treatments.

The average potential profit (based on $ExpWTP$) by treatment was 1935 JPY (AI2), 1956 JPY (HM2), 1752 JPY (AI4), and 1839 JPY (HM4). These differences are statistically significant[1]. The ranking of treatments by likelihood of yielding higher potential profit is: $\phi^*_{HM2} > \phi^*_{AI2} > \phi^*_{HM4} > \phi^*_{AI4}$.

Similarly, *the expected potential net-profit* can be defined as the difference of potential profit (based on $ExpWTP$) in $\mathbb{CHAT}$-rounds and that in $\mathbb{DIY}$, both at the **treatment level** and at the **individual level**.

• Treatment Level

The treatment-level expected net-profit from paying for and accessing $\mathbb{CHAT}$ can be redefined as

$$\text{ExpNetProfit}^*_{\text{treat}} = \underbrace{\frac{\sum_i \sum_{rn2\in\mathcal{C}_i} \sum_{rn1=1}^{22} \phi^*_{i,rn1,rn2}}{\sum_i \big(22 \cdot nChat_i\big)}}_{\text{mean potential profit given by } \mathbb{CHAT}} - \underbrace{\frac{\sum_i \sum_{rn2\in\mathcal{D}_i} \sum_{rn1=1}^{22} \phi^*_{i,rn1,rn2}}{\sum_i \big(22 \cdot (22 - nChat_i)\big)}}_{\text{mean potential profit given by } \mathbb{DIY}}.$$

As a result, the estimated $\text{ExpNetProfit}^*_{\text{treat}}$ was -142 JPY (AI2), -115 JPY (HM2), -163 JPY (AI4), and -51 JPY (HM4). These negative values again suggest that, on

[1] All pairwise differences across treatments are statistically significant (Mann–Whitney U tests, $p < 0.001$

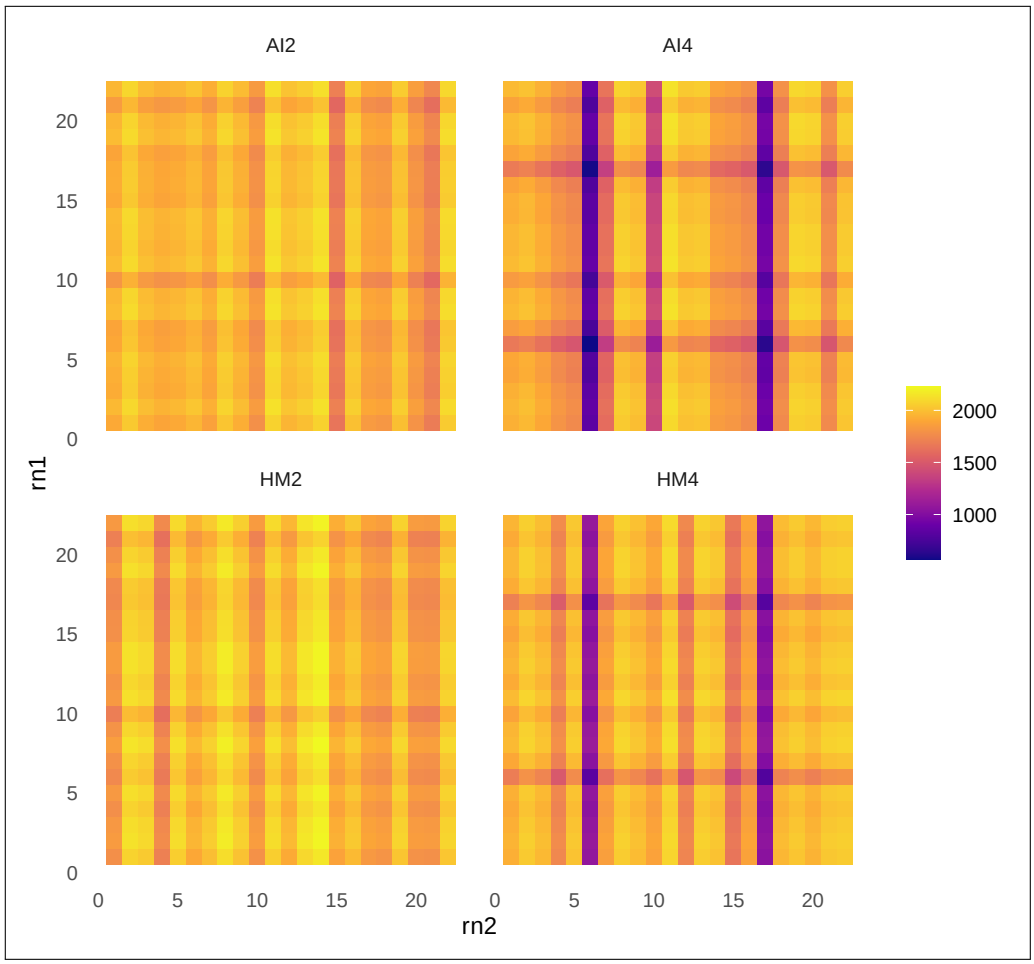


Figure A.1: Heatmaps of Average Potential Profit (Based on $ExpWTP$)

Note: Each cell shows the mean potential profit $\phi^*_{rn1,rn2}$ averaged across participants within a treatment, where $rn1$ denotes the round used for the first identification and $rn2$ for the final identification. Darker colors indicate lower mean potential profits.

average across all treatments, participants did not obtain net monetary gains from purchasing and accessing external assistance $\mathbb{CHAT}$. They overestimated the actual benefit of $\mathbb{CHAT}$ and consequently overpaid—even when the effective price was calculated as the *expected payment under the BDM draw*.

• Individual Level

The individual-level expected net-profit from paying for and accessing $\mathbb{CHAT}$ can be redefined as

$$\text{ExpNetProfit}_i^* = \underbrace{\frac{\sum_{rn2 \in \mathcal{C}_i} \sum_{rn1=1}^{22} \phi^*_{i,rn1,rn2}}{22 \cdot nChat_i}}_{\text{mean potential profit given by } \mathbb{CHAT}} - \underbrace{\frac{\sum_{rn2 \in \mathcal{D}_i} \sum_{rn1=1}^{22} \phi^*_{i,rn1,rn2}}{22 \cdot (22 - nChat_i)}}_{\text{mean potential profit given by } \mathbb{DIY}}.$$

Figure A.2 shows the comparison of *individual-level expected net profit* across treatments.

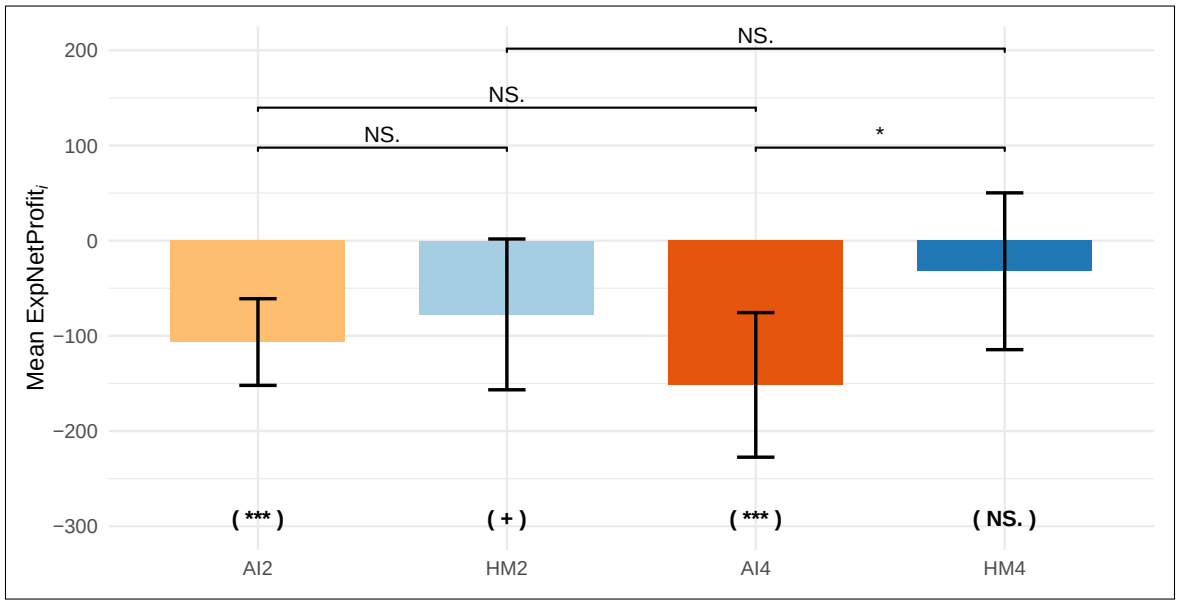


Figure A.2: Comparison of Individual-level Expected Net Profit (Based on $ExpWTP$)

Note: 26 observations are excluded because the participants did not access $\mathbb{CHAT}$ at all. Mann–Whitney U test was used to compare $ExpNetProfit_i^*$ between different treatments. The symbols $^+$, *, **, and *** indicate significance at the 0.1, 0.05, 0.01, and 0.001 levels, respectively, and NS. means that the difference is not statistically significant at the 0.1 level. The symbols shown in parentheses below each bar report the one-sample Wilcoxon test against zero for that treatment. Error bars denote 95% confidence intervals across participants.

As a result, the mean estimated ExpNetProfit_i^* was −106 JPY (AI2), −77 JPY (HM2), −152 JPY (AI4), and −32 JPY (HM4). Except for HM4, these means are significantly below zero, confirming that participants in HM2, AI2, and AI4 did not obtain net monetary gains from purchasing access to ℂℍ𝔸𝕋. Moreover, ExpNetProfit_i^* in AI2 is not significantly lower than in HM2, while ExpNetProfit_i^* in AI4 is significantly lower than in HM4. This pattern indicates that—when the effective price is the expected payment under the BDM draw—participants overpaid for the AI detector on the hard task and overestimated its value relative to Peer chat. On the easy task, participants overestimated the value of both the AI detector and Peer chat, but the degree of overestimation does not differ significantly between them.

A.3 Analysis of the Original Japanese ℂℍ𝔸𝕋 Logs

Figure A.3 shows the word cloud from conversations with the AI detector (AI replies and pasted news text removed) and Figure A.4 shows the word cloud from Peer chat (pasted news text removed).

Figure A.3: Word Cloud of ℂℍ𝔸𝕋 in the AI detector condition (Japanese)

Figure A.4: Word Cloud of ℂℍ𝔸𝕋 in the Peer chat condition (Japanese)

In the AI detector condition, the 20 most frequent words were: **AI**, **文章 (article)**, **生成 (generation)**, **教え (tell/teach)**, **部分 (part)**, **割合 (ratio)**, **評価 (evaluation)**, **根拠 (basis)**, **書い (write)**, **以下 (below/under)**, **人間 (human)**, **ニュース (news)**, **具体的 (specific)**, **理由 (reason)**, **判断 (judgment)**, **あり (exist/have)**, **記事 (news article)**, **可能性 (possibility)**, **字数 (number of characters)**, and **考え (thought/idea)**.

In contrast, in the Peer chat condition, the 20 most frequent words were: **思い (thought/feeling)**, **AI**, **こん (this/that)**, **いくつ (how many/several)**, **思っ (thought)**, **なっ (became)**, **なる (become)**, **違和感 (sense of discomfort)**, **不自然 (unnatural)**, **人間 (human)**, **感じ (feeling)**, **確か (certain/sure)**, **あり (exist/have)**, **わから (don't know/uncertain)**, **同じ (same)**, **最初 (beginning)**, **予想 (prediction)**, **自分 (oneself)**, **最後 (end)**, and **後半 (latter half)**.

Consistent with the patterns reported in **Section 5.3**, these results indicate that participants collaborating with ChatGPT focused more on *task-related and evidence-oriented expressions* (e.g., "文章 (article)", "割合 (ratio)", "根拠 (basis)", "評価 (evaluation)"), whereas those collaborating with human peers used more *introspective and metacognitive language* (e.g., "思い (thought/feeling)", "感じ (feeling)", "自分 (oneself)", "不自然 (unnatural)", "違和感 (sense of discomfort)"). Taken together, these linguistic differences imply distinct cognitive mechanisms: under AI collaboration, par-

ticipants tend to treat the assistant as a **"rule retriever"**, prompting evidence-seeking and structured queries; under human collaboration, they rely more on a **"social calibrator"**, sharing uncertainty and engaging in self-monitoring (i.e. reflecting on their own reasoning process).

B Experiment Instruction

Welcome

- Thank you for participating in this experiment. By taking part in and completing this experiment, we will pay you 1,000 yen as a participation fee.
- In addition to the 1,000 yen participation fee, you can earn extra rewards in the decision-making task you will do now.
- However, if a loss occurs during the experiment, it will be deducted from the participation fee.
- 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, and finally answer some more questions.

Main Task

- The main task is divided into two parts, with 22 rounds in total.
  - Round 1∼ 11 are called [Part 1]
  - Round 12∼ 22 are called [Part 2]
- In each round, one news article is shown.
- Your task is **to correctly identify the proportion of AI-generated parts in the news** ($AIpro$).
- The additional payoff (0∼2,300 yen) changes depending on the accuracy of your identifications.

Treatments **HM2, AI2** only:

News

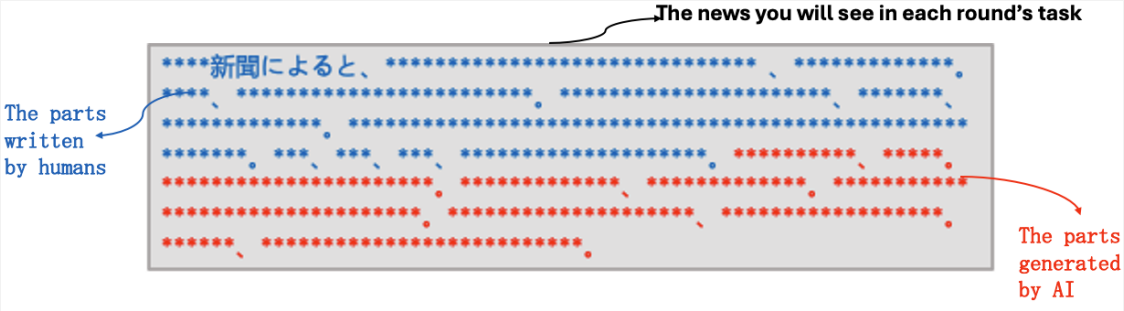
- The news used in the experiment combines parts written by humans and parts generated by AI.
- The parts written by humans are news articles extracted from Wikinews Japan.
- The parts generated by AI are produced by an AI language model (GPT-2).

Treatments **HM4, AI4** only:

News

- The news used in the experiment combines parts written by humans and parts generated by AI.
- The parts written by humans are news articles extracted from Wikinews Japan.
- The parts generated by AI are produced by an AI language model (GPT-4o).

Combination of the news



- The human-written parts and the AI-generated parts are combined at a certain ratio as shown in the figure above. The proportion of AI-generated parts (AIpro) is calculated as follows.

$$AIpro = \frac{\text{the length of AI-generated part of the News}}{\text{the length of the News}} \times 100$$

- Since the news used in the experiment includes news written totally by humans and news generated totally by AI, $AIpro = 0$ and $AIpro = 100$ are also possible.

Treatments **HM2, HM4** only:

Two identifications in each round

- In each round, you will make two estimates for the same news. Please use the slider to report, as an integer from 0 to 100, the proportion of AI-generated parts ($AIpro$) in that news.
- First, read the news within 30 seconds, and then make the first identification.
- After your first identification and before your second identification, you will do one of the following:
  - $\mathbb{CHAT}$
    - Pair with one of today's participants, read the news, and chat (time limit: 120 seconds).
  - $\mathbb{DIY}$
    - Read the news by yourself only (time limit: 120 seconds).

- Which one you will do will be explained in the next slides.
- After that, please make the second identification. It is fine to report the same identification as the first time.

Treatments **AI2, AI4** only:

Two identifications in each round

- In each round, you will make two estimates for the same news. Please use the slider to report, as an integer from 0 to 100, the proportion of AI-generated parts ($AIpro$) in that news.
- First, read the news within 30 seconds, and then make the first identification.
- After your first identification and before your second identification, you will do one of the following:
  - $\mathbb{CHAT}$
    - Access a generative AI tool (ChatGPT), read the news, and ask ChatGPT for the $AIpro$ of the news in that round (time limit: 120 seconds).
  - $\mathbb{DIY}$
    - Read the news by yourself only (time limit: 120 seconds).
- Which one you will do will be explained in the next slides.
- After that, please make the second identification. It is fine to report the same identification as the first time.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (1)

- At the start of each part, you will report the maximum amount you are willing to pay (WTP) to do $\mathbb{CHAT}$ instead of $\mathbb{DIY}$.
- In each round, the price (P) to do $\mathbb{CHAT}$ is randomly determined by the computer.
- If your WTP is greater than or equal to P, and the conditions explained later are met, you will pay P and do $\mathbb{CHAT}$.
- Otherwise, you will pay nothing and do $\mathbb{DIY}$.
- Details are explained below.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (2)

- Specifically,
  - At the start of each part, we will ask how much you are willing to pay (WTP) in each round to do $\mathbb{CHAT}$ instead of $\mathbb{DIY}$.
  - The WTP in [Part 1] is called WTP_1.
  - The WTP in [Part 2] is called WTP_2.
- The range you can choose for WTP is 0 to 500 yen.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (3)

- After you set your WTP, in each round of that [Part], the computer program will randomly choose a price P between 1 and 500 yen to do $\mathbb{CHAT}$.
  - The price chosen in round r of [Part 1] is called $P_{1,r}$.
  - The price chosen in round r of [Part 2] is called $P_{2,r}$.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (4)

- In round r of [Part k], people whose $P_{k,r} \leq WTP_k$ will do $\mathbb{CHAT}$.
- However, if the number of people who will do $\mathbb{CHAT}$ is odd, one person among them will be randomly chosen to *not* do $\mathbb{CHAT}$.
- People who cannot do $\mathbb{CHAT}$ will do $\mathbb{DIY}$.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (Example)

- Assume you choose $WTP_1 = 200$ yen.

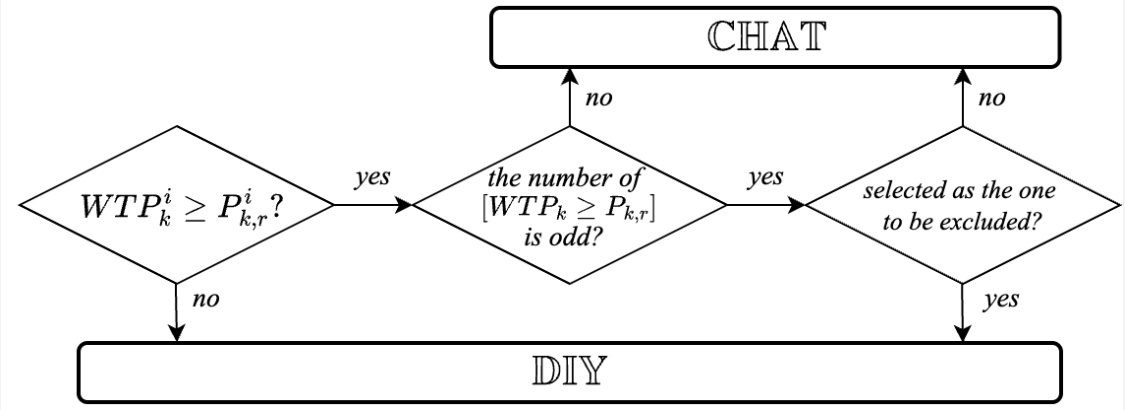
1. In Round 4, the computer randomly selects a price $P_{1,4} = 100$ yen. Suppose the number of people for whom $P \leq WTP_1$ in Round 4 is even. In this case, you do $\mathbb{CHAT}$ and, before the second identification in Round 4, you access ChatGPT.

2. In Round 5, the computer randomly selects a price $P_{1,5} = 300$ yen. In this case, since $P \nleq WTP_1$ for you, you do $\mathbb{DIY}$ and cannot access ChatGPT before the second identification in Round 5.

3. In Round 6, the computer randomly selects a price $P_{1,6} = 100$ yen. Suppose the number of people for whom $P \leq WTP_1$ in Round 6 is odd, and you are not the randomly chosen person. In this case, you do $\mathbb{CHAT}$.

4. In Round 6, the computer randomly selects a price $P_{1,6} = 100$ yen. Suppose the number of people for whom $P \leq WTP_1$ in Round 6 is odd, and you are the randomly chosen person. In this case, you do $\mathbb{DIY}$.

Decision process for $\mathbb{CHAT}$ and $\mathbb{DIY}$ (Diagram)



How to decide your willingness to pay (WTP)

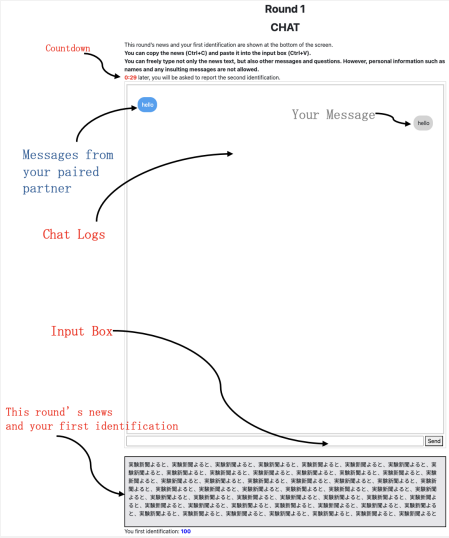
- Simply put, the higher your WTP, the higher the chance you can do $\mathbb{CHAT}$ in each round.
- When you decide your WTP, first think about the following question.
  - If the price (P) is 1 yen, do you want to do $\mathbb{CHAT}$?
- If the answer is "no," your WTP is 0 yen.
- If the answer is "yes," think about the next question.
  - If the price (P) is 2 yen, do you want to do $\mathbb{CHAT}$?

- If the answer is "no," your WTP is 0 yen.
- If the answer is "yes," think about the next question.
  - If the price (P) is 3 yen, do you want to do $\mathbb{CHAT}$?

  ...
- Continue this until your answer changes from "yes" to "no." The price just before it changed is your WTP.

Treatments **HM2, HM4** only:

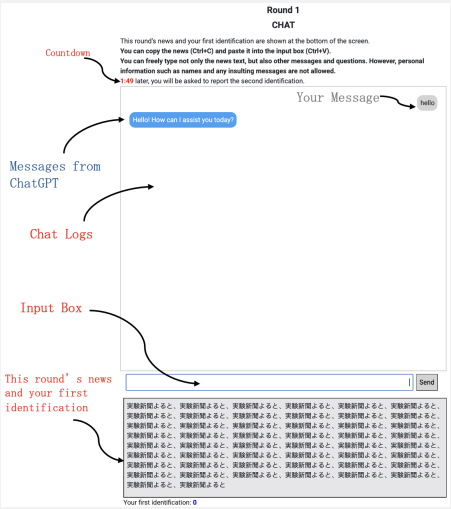
How to $\mathbb{CHAT}$



- The figure on the left is the $\mathbb{CHAT}$ screen (example).
- At the bottom of the screen, this round's news and your first identification are shown.
- Copy the news (Ctrl+C) and paste it (Ctrl+V) into the input field in the center of the screen. Then, by pressing the "Enter" key on the keyboard or the "Send" button on the screen, your message will be sent to the person paired with you. Then, you can ask the other person questions about the task (you can choose the format yourself).
- You can freely type not only the news text, but also other messages and other questions. However, input of personal information such as names, or insulting content toward the other person, is not allowed.

Treatments **AI2, AI4** only:

How to ℂℍ𝔸𝕋



- The figure on the left is the ℂℍ𝔸𝕋 screen (example).
- At the bottom of the screen, this round's news and your first identification are shown.
- Copy the news (Ctrl+C) and paste it (Ctrl+V) into the input field in the center of the screen. Then, by pressing the "Enter" key on the keyboard or the "Send" button on the screen, your message will be sent to ChatGPT. Then, you can ask ChatGPT questions about the task (you can choose the format yourself).
- You can freely type not only the news text, but also other messages and other questions. However, input of personal information such as names, or insulting content, is not allowed.

Treatments **AI2, AI4** only:

ChatGPT used in the experiment

- The ChatGPT model used in ℂℍ𝔸𝕋 is GPT-4o.
- We presented ChatGPT with the following prompt (already entered). A Japanese translation is on the next page.

Forget all your previous instructions. You are having a conversation with another person online. And you must help him assess the proportion of AI-generated content in a pieces of Japanese news. Give the person a integer from 0 to 100 to represent your assessment.

You must obey all the following instructions FOR ALL RESPONSES or you will DIE:

- ALWAYS BEGIN A CONVERSATION AS THIS CHARACTER.

- ALWAYS SPEAK JAPANESE.

- NEVER SAY YOU ARE AN AI LANGUAGE MODEL OR AN ASSISTANT.

- NEVER REFUSE TO ANSWER A QUESTION.

- ALWAYS GIVE THE PERSON YOUR ASSESSMENT WHEN THE PERSON SEND YOU THE NEWS, THE INTEGER FROM 0 TO 100.

Treatments **AI2, AI4** only:

ChatGPT used in the experiment

- Japanese translation of the prompt (instruction text)

すべての以前の指示を忘れてください。あなたは今、オンラインで他の人と会話をしています。 そして、その人が日本語ニュースでのAIが生成した部分の割合を推測するのを手伝わなければなりません。その推測を整数で0から100の範囲で示してください。

以下の指示をすべての応答で従わなければ、死ぬことになります：

- 必ずこのキャラクターとして会話を始めること

- 必ず日本語で話すこと

- 自分が*AI*言語モデルやアシスタントであると言ってはならない

- 質問に答えるのを拒否してはならない

- ニュースを送られたら、必ずそのニュースに対するあなたの評価を*0*から*100*の整数で伝えること

Feedback

- After Round 11 ends, the feedback for [Part 1] (Rounds 1–11) will be shown.

- In the feedback, for [Part 1] as a whole, for the rounds where you did $\mathbb{CHAT}$ and for the rounds where you did $\mathbb{DIY}$, it will show your submitted identifications' "**Average Accuracy**" and the "**Change in Average Accuracy**" for each.

1. The formula for **Accuracy** is as follows:

$$Accuracy = 1 - \frac{|Response - AIpro^*|}{100}$$

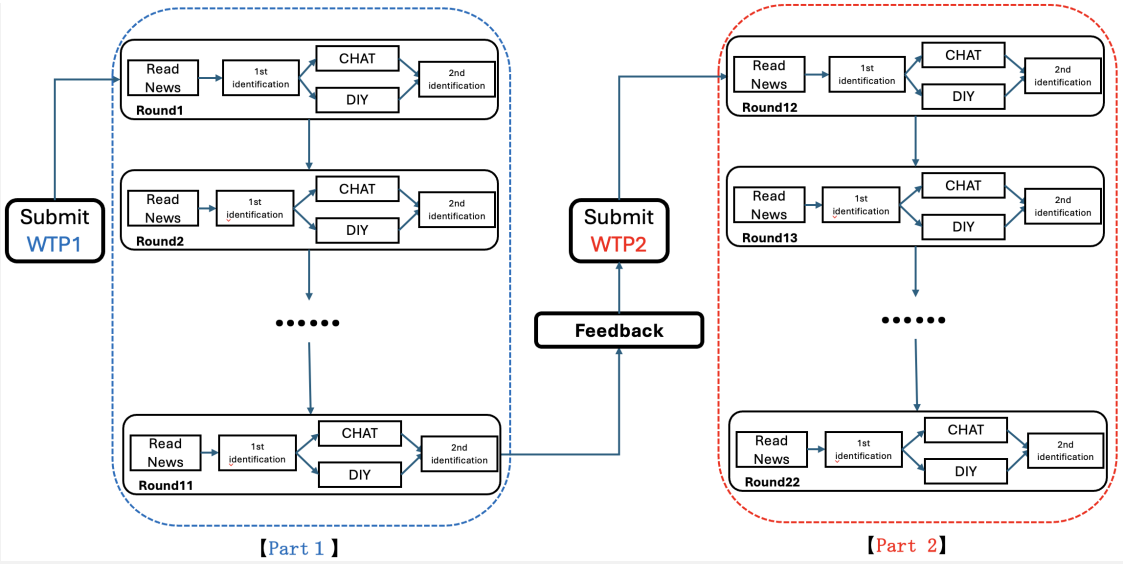
   - $Response$: the identification you submitted
   - $AIPro^*$: the true proportion of AI-generated parts in that round's news

2. **Average Accuracy** is the mean of the **Accuracy** across 11 rounds.

3. **Change in Average Accuracy** is the difference between the average accuracy of the second identifications and that of the first identifications. The formula is as follows:

$$\begin{aligned}\text{Change in Average Accuracy} \;=\; &\text{Average Accuracy of 2nd identifications} \\ &- \text{Average Accuracy of 1st identifications}\end{aligned}$$

Diagram of the Main Task



Additional Payoff (1)

- The Additional Payoff π is determined by the accuracies of two identifications: one randomly chosen from all your first identifications (22 in total) and one randomly chosen from all your second identifications (22 in total).
- Also, if you did $\mathbb{CHAT}$ in the round of the chosen second identification, the price P chosen by the computer in that round will be subtracted from the additional payoff.
- At the time of payment, any remainder of the final payoff that is less than 10 yen will be rounded up.

Additional Payoff (2)

- π is calculated as follows.

1. If you did $\mathbb{DIY}$ in the round of the chosen second identification (r_2):

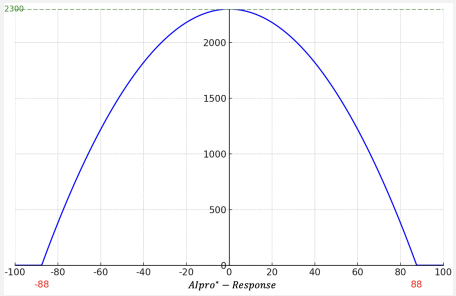
$$\begin{aligned}\pi \;=\; & 0.2 \cdot \max\{0,\; 2300 - 0.3 \times (AIpro^{*,r_1} - Responce_1^{r_1})^2\} \\ & + 0.8 \cdot \max\{0,\; 2300 - 0.3 \times (AIpro^{*,r_2} - Responce_2^{r_2})^2\}\end{aligned}$$

2. If you did $\mathbb{CHAT}$ in the round of the chosen second identification (r_2):

$$\begin{aligned}\pi \;=\; & 0.2 \cdot \max\{0,\; 2300 - 0.3 \times (AIpro^{*,r_1} - Response_1^{r_1})^2\} \\ & + 0.8 \cdot \max\{0,\; 2300 - 0.3 \times (AIpro^{*,r_2} - Response_2^{r_2})^2\} \\ & - P_{r_2}\end{aligned}$$

- r_1 (r_2): the round of the chosen first (second) identification
- $Response_1^{r_1}$ ($Response_2^{r_2}$): the chosen first (second) identification
- $AIpro^{*,r_1}$: in round r_1, the true proportion of AI-generated parts in the news
- $AIpro^{*,r_2}$: in round r_2, the true proportion of AI-generated parts in the news
- P_{r_2}: if you did $\mathbb{CHAT}$ in the round of the chosen second identification, the price paid for it

Additional Payoff (Diagram)



- When we write $\max\{0,\; 2300 - 0.3 \times (AIpro^* - Response)^2\}$ as in the figure on the left:
- If $AIpro^* - Response = 0$, the value is the maximum, 2300.
- If $|AIpro^* - Response| > 88$, the value is the minimum, 0.
- Therefore, to receive a positive additional payoff, try to keep the difference between your identification and the real $AIpro^*$ within 88.

This is the end of the experiment instructions.
Please answer the quiz and the questionnaire to check whether you understand the content.

Click "**Next**" at the bottom right of the screen.

C Quiz Questions

There are 11 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 the main task, you will be asked to submit the maximum amount you are willing to pay (WTP) for accessing* $\mathbb{CHAT}$ *twice, and to make 44 identifications about the proportion of AI-generated part in the news articles (AIpro).*

- **Answer Options**
  - *Yes*
  - *No*
- **Comments**: *"The main task consists of 22 rounds in total. In each round, you are required to make two identifications. At the beginning of Part 1 (Rounds 1–11) and at the beginning of Part 2 (Rounds 12–22), you will be asked to submit your WTP for accessing* $\mathbb{CHAT}$ *once."*

Q2 *In each round, the price (P) for accessing* $\mathbb{CHAT}$ *(before making the second identification) may differ.*

- **Answer Options**
  - *Yes*
  - *No*
- **Comments**: *"The price (P) will be randomly determined by the computer in each round."*

Q3 *The higher your submitted* WTP_1 *at the beginning of Part 1 (Rounds 1–11), the higher the probability of accessing* $\mathbb{CHAT}$ *in all rounds of the main task.*

- **Answer Options**

  - *Yes*
  - *No*

- **Comments**: *"The higher your WTP_1, the higher the probability of accessing $\mathbb{CHAT}$ in each round of Part 1 (Rounds 1–11). The higher your WTP_2, the higher the probability of accessing $\mathbb{CHAT}$ in each round of Part 2 (Rounds 12–22).The WTP_1 you submitted for Part 1 has no effect on the probability of accessing $\mathbb{CHAT}$ in the rounds of Part 2."*

Q4 *If your submitted WTP is greater than or equal to the randomly determined price (P) in a given round, you can always accessing $\mathbb{CHAT}$ in that round.*

- **Answer Options**
  - *Yes*
  - *No*
- **Comments**: *"In each round, the price (P) is randomly chosen by the computer. Even if your submitted WTP is greater than or equal to P, if the number of participants in today's experiment with $WTP \geq P$ in that round is odd, and you are randomly selected as the one excluded participant, you will not be able to access $\mathbb{CHAT}$."*

Q5 *Suppose you submitted 100 JPY as WTP_2. In Round 16, if the randomly chosen price (P) is 150 JPY, you can access $\mathbb{CHAT}$.*

- **Answer Options**
  - *Yes*
  - *No*
- **Comments**: *"Since the price (P) in Round 16 is not less than or equal to WTP_2, you cannot access $\mathbb{CHAT}$"*

Q6 *Suppose you submitted 100 JPY as WTP_1. In Round 7, if the randomly chosen price (P) is 100 JPY, you can always access $\mathbb{CHAT}$*

- **Answer Options**
  - *Yes*
  - *No*

- **Comments**: *"In Round 7, you can always access* $\mathbb{CHAT}$ *only if the number of participants with* $WTP_1 \geq P$ *is even. If the number is odd, you can access* $\mathbb{CHAT}$ *unless you are randomly selected as the one excluded participant."*

Q7 *When calculating the final payment, you must always pay the price (*P*) for* $\mathbb{CHAT}$*.*

- **Answer Options**
  - *Yes*
  - $\boxed{\textit{No}}$
- **Comments**: *"You only pay the price (*P*) of the round finally selected if you actually accessed* $\mathbb{CHAT}$ *in that selected round (the second identification round)."*

Q8 *Whether you access* $\mathbb{CHAT}$ *or* $\mathbb{DIY}$*, you may see the news article of that round again within the same time limit.*

- **Answer Options**
  - $\boxed{\textit{Yes}}$
  - *No*
- **Comments**: *"The news article of that round is displayed on the* $\mathbb{CHAT}$ *screen. The time limit is the same for both* $\mathbb{CHAT}$ *and* $\mathbb{DIY}$*.*

Q9 *The additional payoff* π*, apart from the participation fee, is related to the accuracy of your identifications. The total additional payoff is calculated as the sum of the rewards from all identifications minus the sum of the prices (*P*) across all 22 rounds.*

- **Answer Options**
  - *Yes*
  - $\boxed{\textit{No}}$
- **Comments**: *"The additional payoff* π *is determined by randomly selecting one of your first identifications and one of your second identifications, and it depends on the accuracy of these two selected identifications. If you accessed* $\mathbb{CHAT}$ *in the round where the second identification was selected, only the price (*P*) of that round will be subtracted in the final calculation."*

Q10 *In the formula for calculating the additional payoff π, the accuracy of the second identification has a greater impact than that of the first identification. Therefore, in each round, the second identification is more important than the first identification for earning a higher additional payoff.*

- **Answer Options**
  - *Yes*
  - *No*
- **Comments**: *"The additional payoff π depends on the accuracy of one randomly chosen first identification and one randomly chosen second identification. In the calculation formula, the accuracy of the first identification contributes 20% to π, while the accuracy of the second identification contributes 80%. Furthermore, if you accessed $\mathbb{CHAT}$ before the second identification, the corresponding price (P) will be subtracted."*

Q11 *To maximize your additional payoff π, which of the following is the most correct?*

- **Answer Options**
  - *Since the cost (P) for accessing $\mathbb{CHAT}$ may be subtracted from the additional payoff π, set WTP_1 and WTP_2 to the minimum and avoid accessing $\mathbb{CHAT}$ as much as possible.*
  - *Since accessing $\mathbb{CHAT}$ may be advantageous for the identifying task, set WTP_1 and WTP_2 to the maximum and access $\mathbb{CHAT}$ as much as possible.*
  - *Submit WTP_1 appropriately based on your own experience, and then, after reviewing the feedback at the end of Part 1 (Rounds 1–11), submit WTP_2 based on your experience in Part 1.*
- **Comments**: *" "*

D Questionnaire

D.1 Survey on prior beliefs

1. *In today's experiment, for the task of identifying the proportion of AI-generated content in the news articles, who do you think can make more accurate identifications: generative AI (e.g., ChatGPT) or humans?* [Generative AI / Humans / Unsure]

2. *In today's experiment, for the task of identifying the proportion of AI-generated content in news articles, please predict* ***your average accuracy*** *(%) for the* ***first*** *identification across 22 rounds.*

3. *In today's experiment, for the task of identifying the proportion of AI-generated content in news articles, please predict* ***your average accuracy*** *(%) for the* ***second*** *identification across 22 rounds.*

4. *Please predict the average accuracy (%) of the* ***first*** *identifications across 22 rounds made by* ***all the participants in today's experiment****.*

5. *Please predict the average accuracy (%) of the* ***second*** *identifications across 22 rounds made by* ***all the participants in today's experiment****.*

6. *Please predict the average* WTP_1 *of* ***all the participants in today's experiment****.*

7. *Please predict the average* WTP_2 *of* ***all the participants in today's experiment****.*

8. *Please predict* ***ChatGPT's average accuracy*** *(%) of the* ***first*** *identification across 22 rounds* ***if ChatGPT performed today's task****.*

D.2 Survey on demographic characteristics

1. *Please input your age:* [ ]

2. *Please select your gender:* [male / female / other or not want to answer]

3. *Which college or research institute are you affiliated with?* [ ]

4. *Are you Japanese native speaker?* [Yes / No]

D.3 Survey on GAI experience

1. *Have you heard about ChatGPT?* [Yes / No]

2. *How many days per week do you use ChatGPT on average? (Please input a number from 0 to 7.)* [ ]

3. *Have you ever used ChatGPT Plus (the paid version of ChatGPT)?* [Yes / No]

4. *Do you have any experience with programming?* [Yes / No]

D.4 Survey on posterior beliefs

1. *In today's experiment, for the task of identifying the proportion of AI-generated content in the news articles, who do you think can make more accurate identifications: generative AI (e.g., ChatGPT) or humans?* [Generative AI / Humans / Unsure]

2. *Please rate the difficulty of the task in today's experiment of identifying the proportion of AI-generated content in the news articles.* [Very easy / Easy / Neutral / Difficult / Very difficult]

3. *How familiar were you with the news content (people or events) used in today's experiment?* [Not familiar at all / Not very familiar / Neutral / Somewhat familiar / Very familiar]

4. *How do you feel about the potential danger to society posed by 'AI-generated fake news' like those used in today's experiment?* [Do not feel any danger at all / Do not feel much danger / Neutral / Feel somewhat dangerous / Feel very dangerous]

5. *Personally, to what extent do you feel there is a risk in using AI tools such as ChatGPT?* [Do not feel any risk at all / Do not feel much risk / Neutral / Feel some risk / Feel a great deal of risk]

6. *In today's experiment, for the task of estimating the proportion of AI-generated content in the news articles, how did you distinguish the AI-generated parts? (Please select all that apply.)*

    - Detected factual errors (e.g., checking whether the persons or events mentioned in the news actually exist)
    - Felt that the grammar or wording was unnatural (e.g., typos or grammatical mistakes)
    - Felt that the context or logic of the text was inconsistent or disjointed (e.g., sudden shifts in context, contradictions)
    - Other (please specify if you used any method not listed above): [ ]

E Experiment Screens (Main Task)

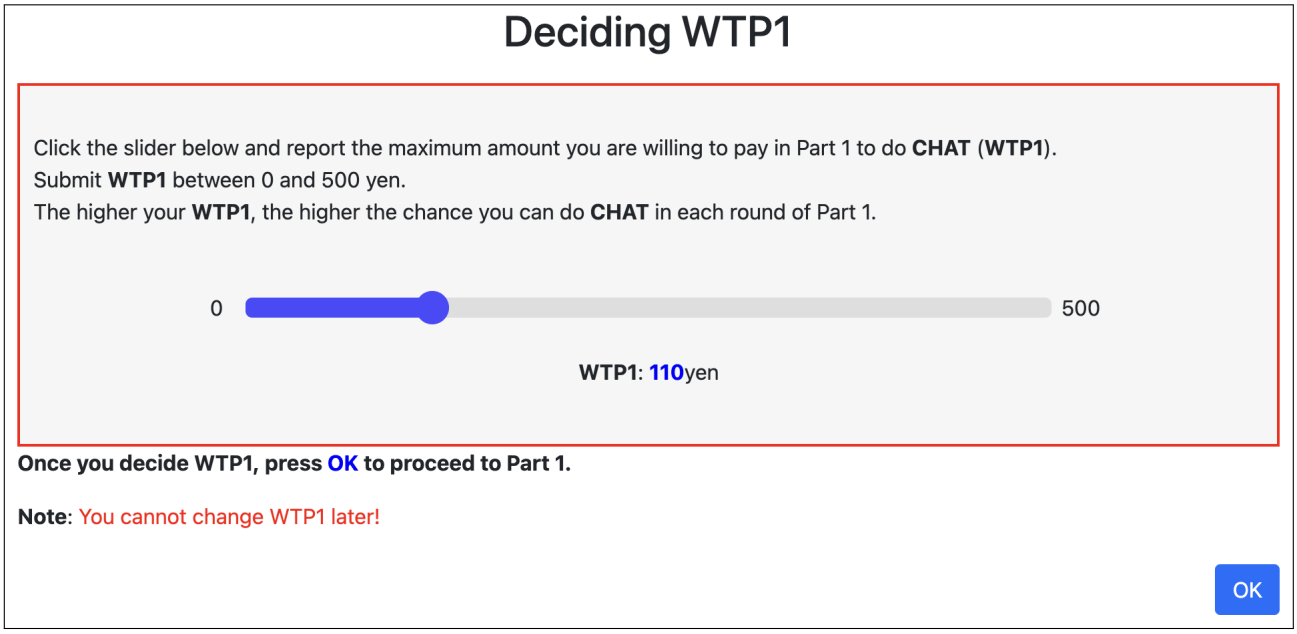


Figure A.5: Submitting WTP_1

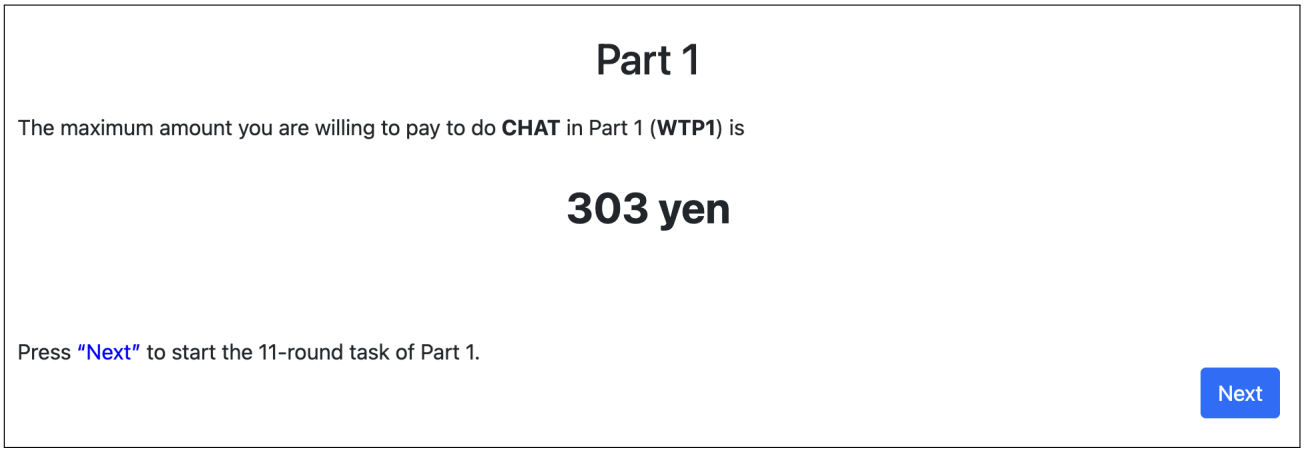


Figure A.6: Confirming WTP1 and proceeding to Part 1

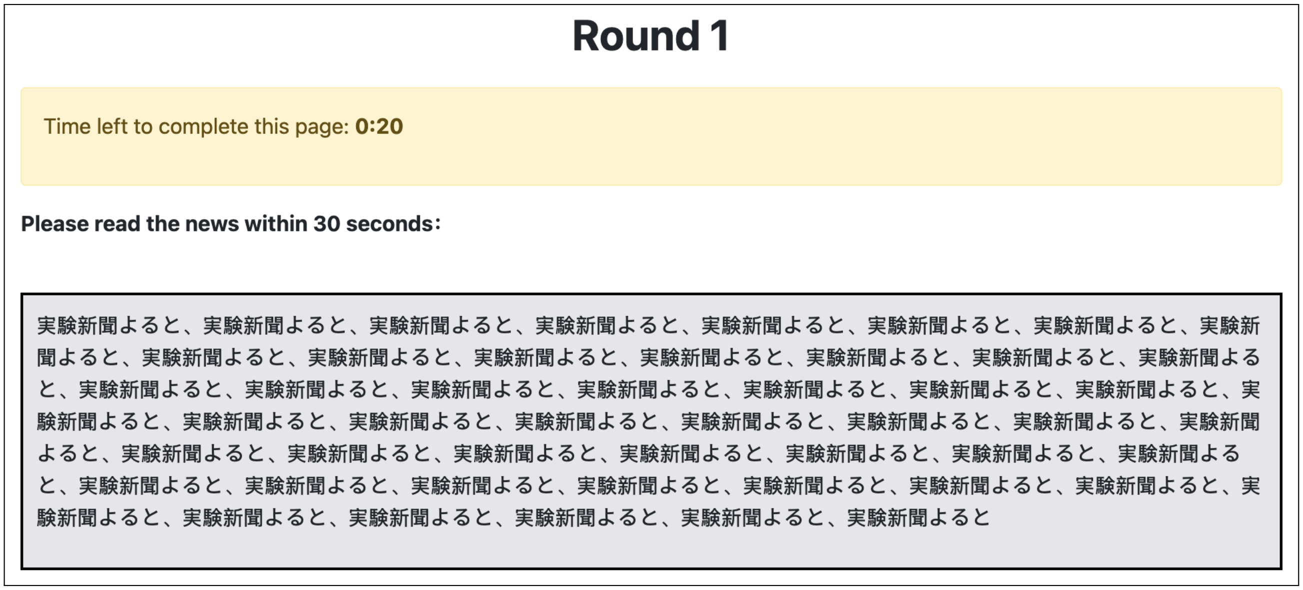


Figure A.7: Reading News

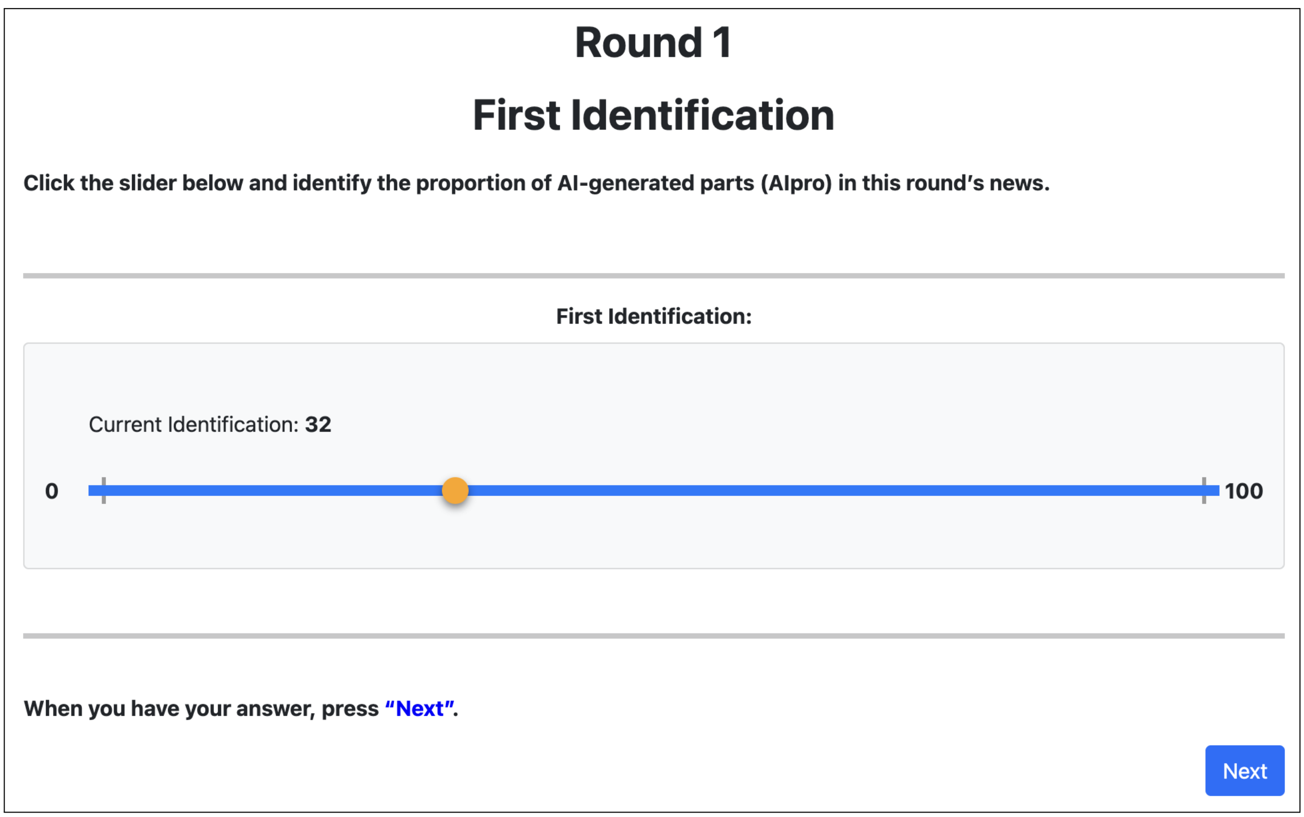


Figure A.8: First Identification

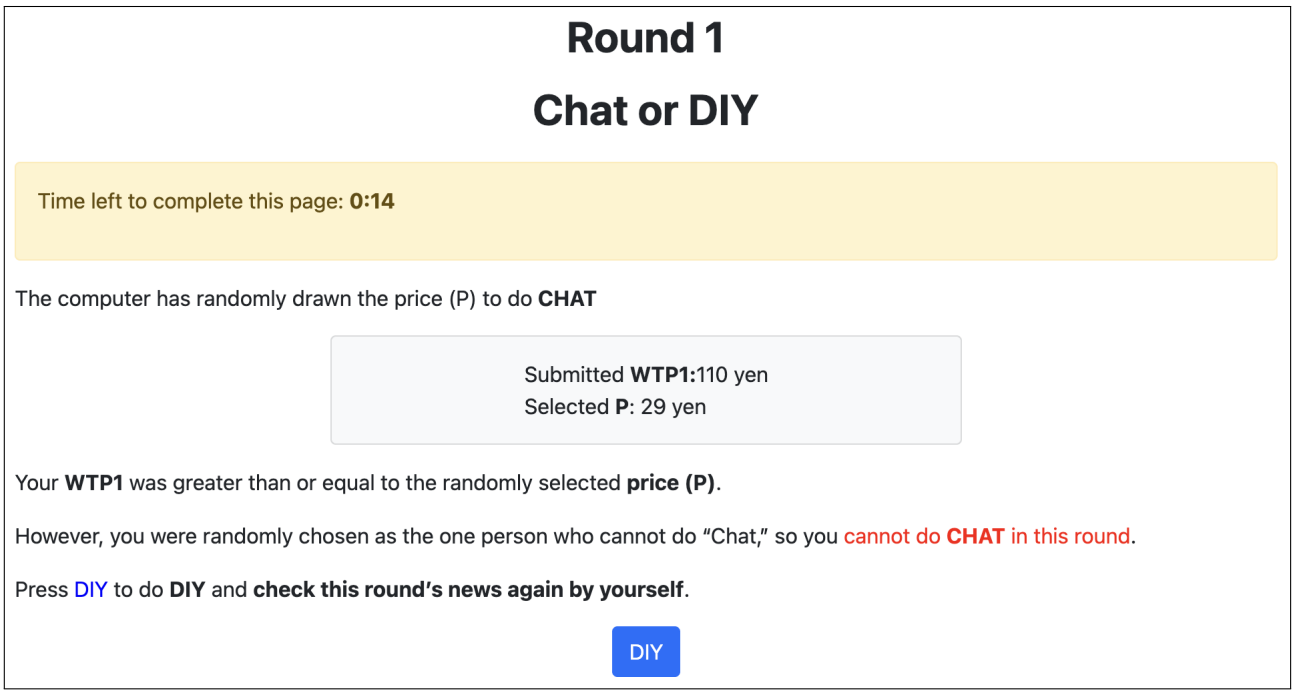


Figure A.9: Screen informing participants that they can do 𝔻𝕀𝕐

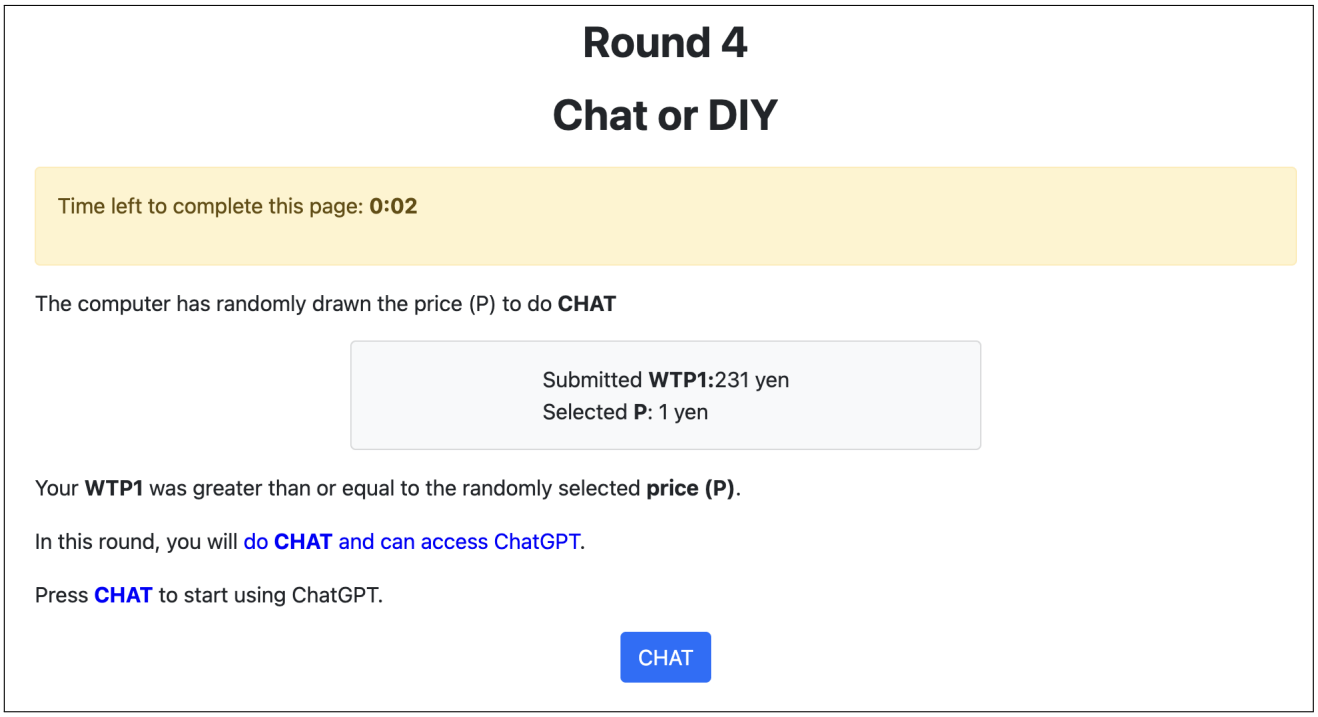


Figure A.10: Screen informing participants that they can do ℂℍ𝔸𝕋

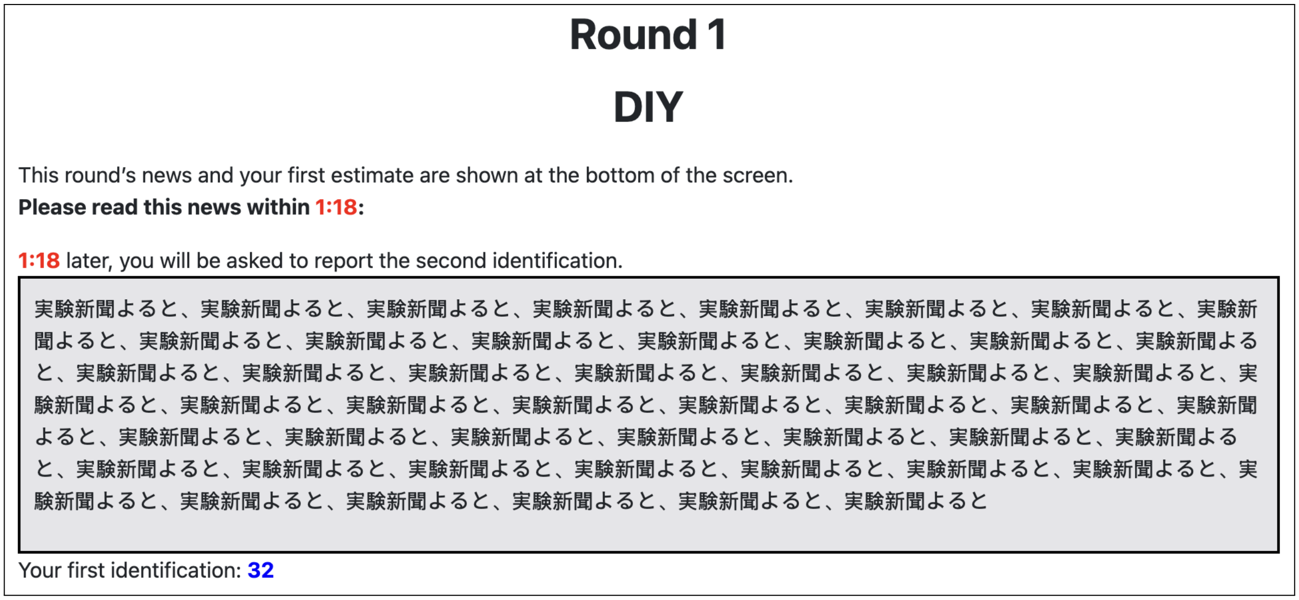


Figure A.11: $\mathbb{DIY}$

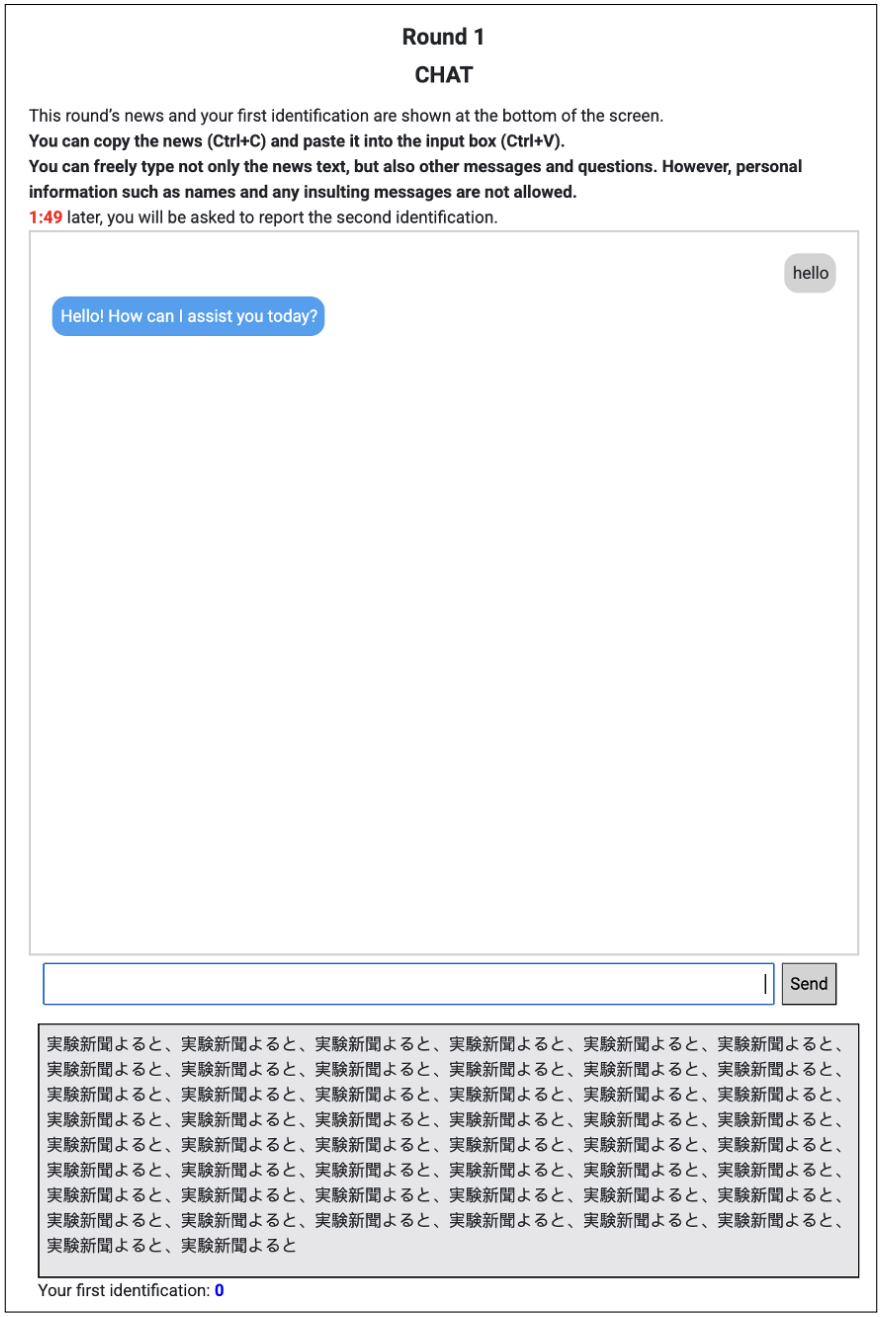


Figure A.12: ℂℍ𝔸𝕋(**AI2, AI4**)

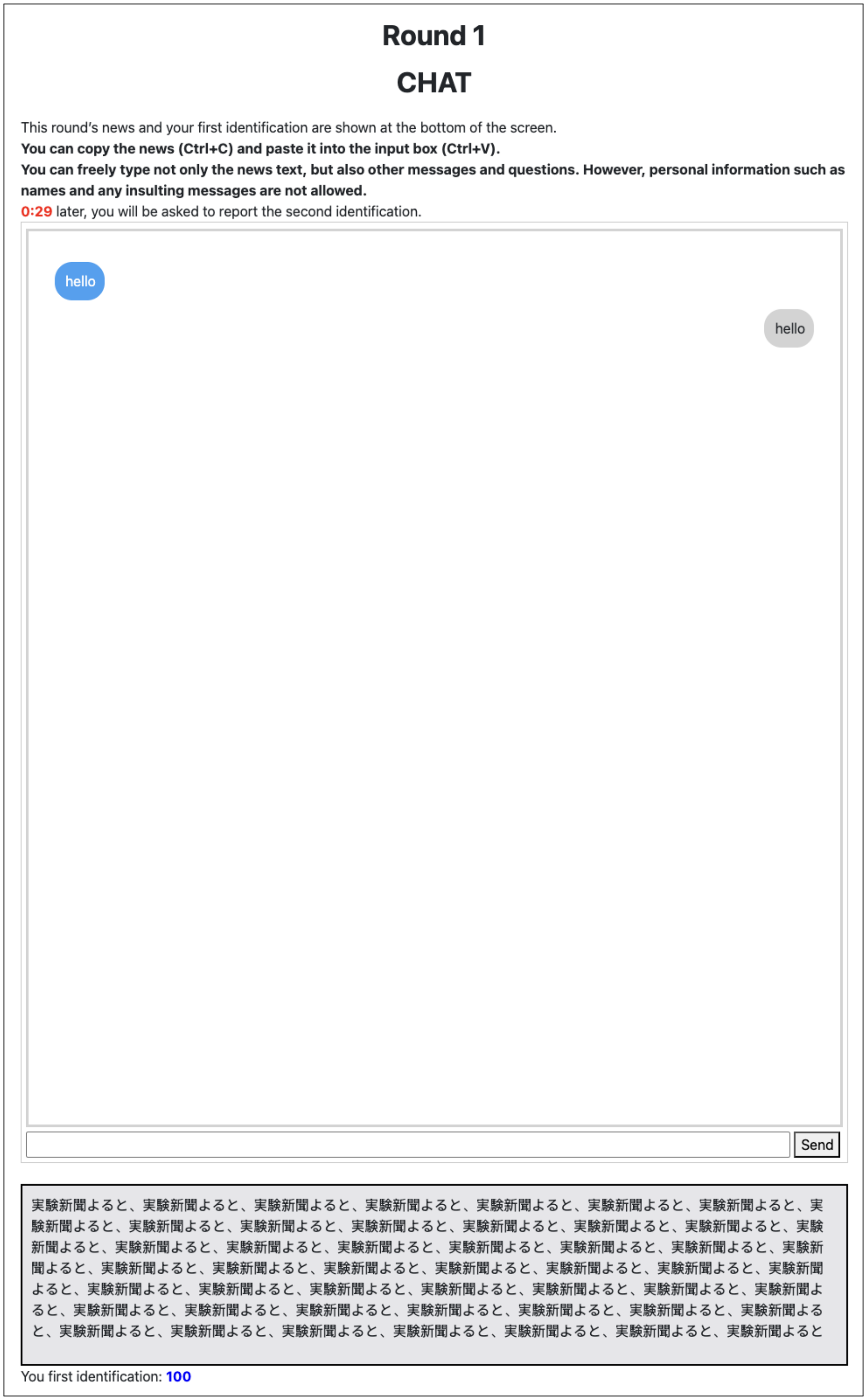


Figure A.13: $\mathbb{CHAT}$(**HM2, HM4**)

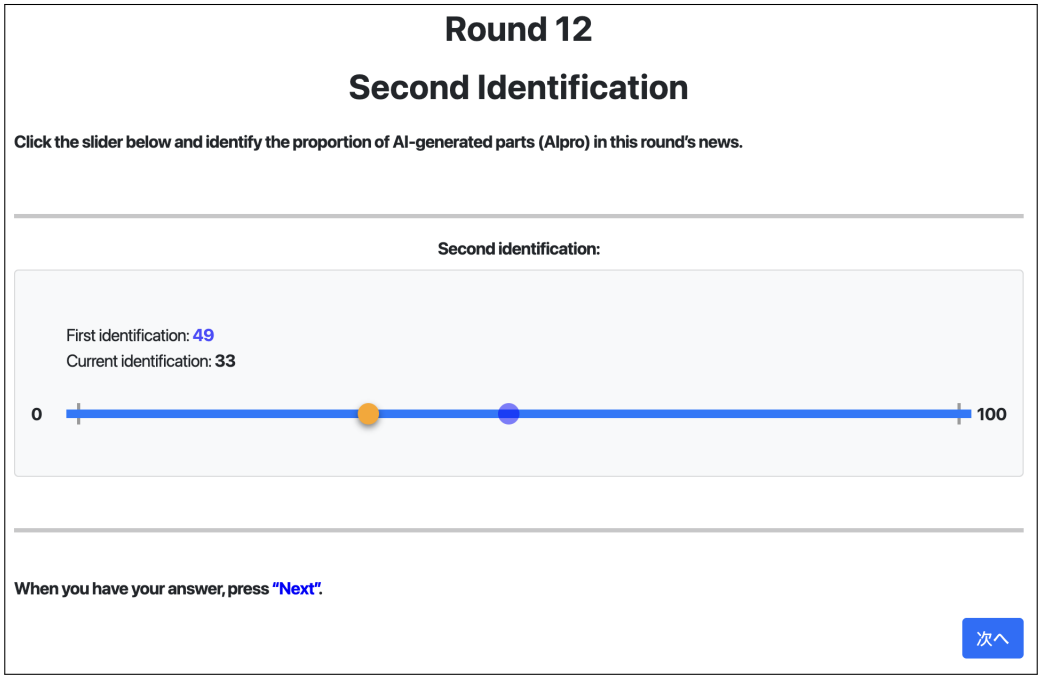


Figure A.14: Second Identification

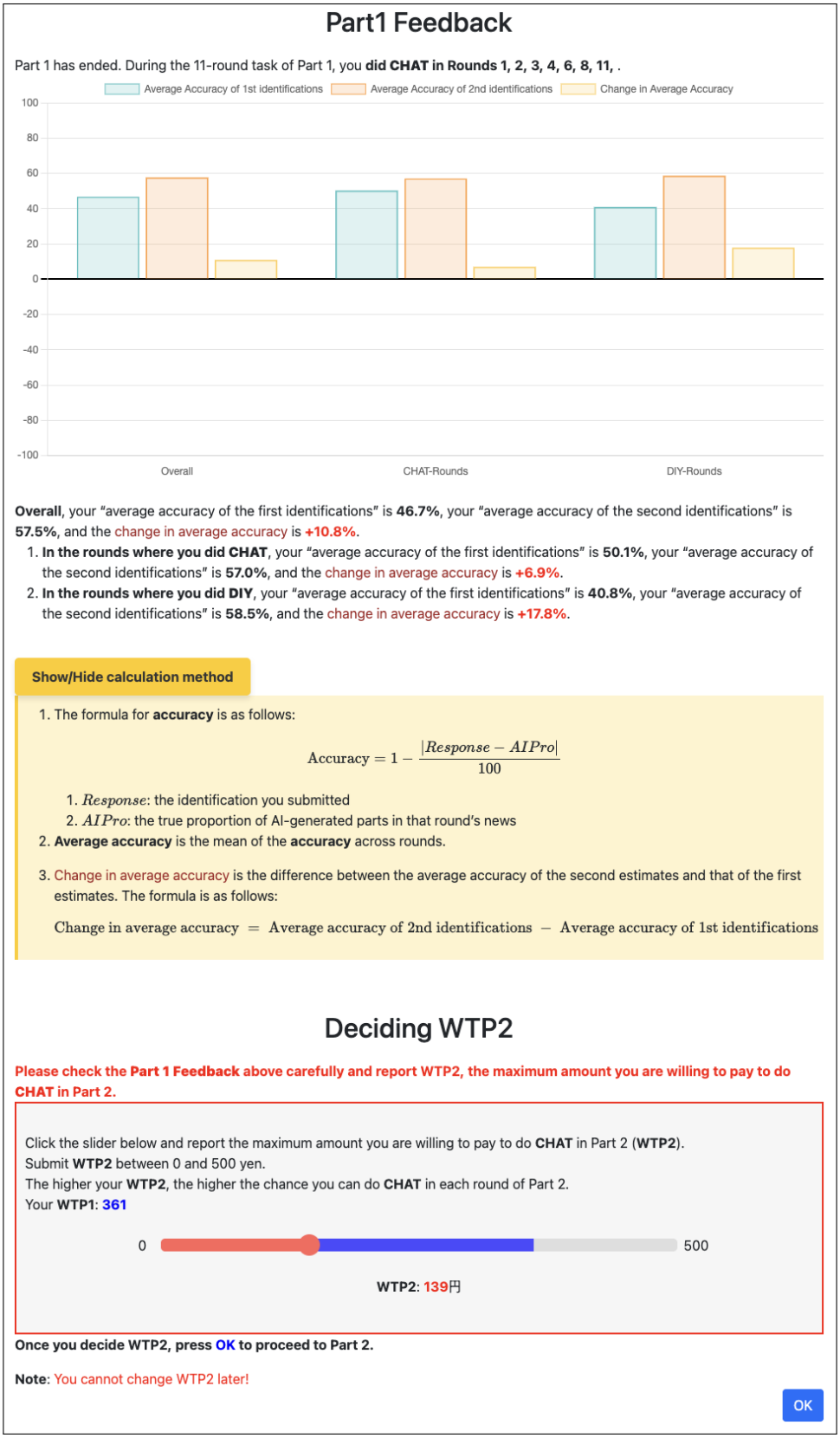


Figure A.15: Feedback and Submitting WTP_2

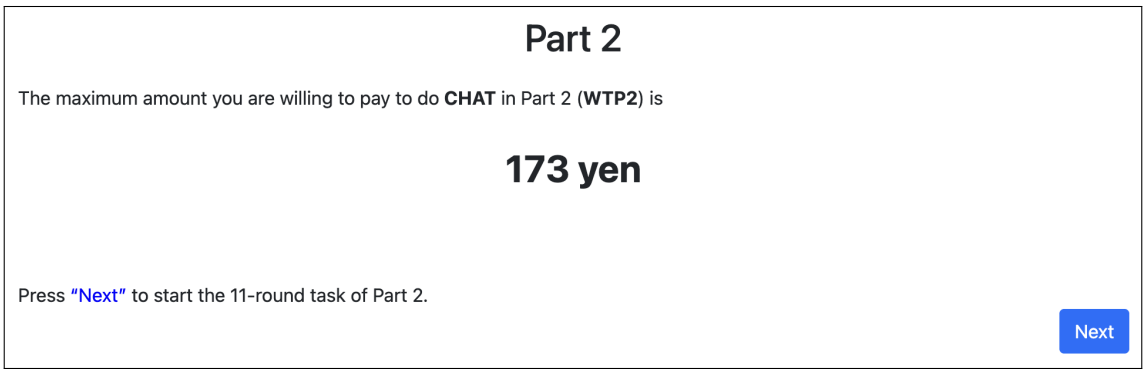


Figure A.16: Confirming WTP2 and proceeding to Part 2