Paying AI to detect AI[*]

Yuhao Fu[a] **Nobuyuki Hanaki**[b]

Abstract

We embedded a ChatGPT-based AI detector into a laboratory setting to investigate whether participants are willing to pay more to collaborate with AI rather than with human peers in detecting deepfake (AI-generated) news. In multiple rounds of deepfake detection tasks, student participants were incentivized to assess the proportion of AI-generated content in the deepfake news, with task difficulty varying depending on the large language model (GPT-2 vs. GPT-4o) used to generate the content. We found that participants demonstrated a higher willingness-to-pay (WTP) for the AI detector compared to human peers to detect AI, despite the AI detector not providing superior assistance. Although GPT-4o–generated news proved more difficult to detect than GPT-2–generated news, participants' WTP for external assistance did not increase when facing more challenging tasks. Our study reveals an over-reliance on AI and raises concerns about the spread of deepfakes, thereby contributing to a deeper understanding of human-AI interaction and supporting advancements in deepfake detection in the Generative Artificial Intelligence (GAI) era.

Keywords: ChatGPT, Human-AI collaboration, willingness to pay, AI reliance, deepfake detection

JEL classification: C90; D83; D90;

[*] This research has benefited from the financial support of (a) the Joint Usage/Research Center, the Institute of Social and Economic Research (ISER), and the University of Osaka, and (b) Grants-in-aid for Scientific Research No. 20H05631 and 23H00055 from the Japan Society for the Promotion of Science. The design of the experiment reported in this paper was approved by the Research Ethics Committee at the Institute of Social and Economic Research, the University of Osaka (#20240904) and the Research Ethics Committee at the Research Institute for Socionetwork Strategies, Kansai University (#2024030). The experiment is preregistered at aspredicted.org (#208604).

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

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

1. Introduction

The rapid development of generative AI has brought new challenges to human decision-making. One of the most pressing issues is the spread of AI-generated misinformation, such as "deepfake" news. Detecting such content is difficult for individuals, which motivates the demand for detection tools. In recent years, ChatGPT-based AI detectors have been introduced to assist users in detecting deepfakes. However, "AI reliance" -- the tendency to over-trust AI tools and to privilege machine output -- has also become a concern (Passi and Vorvoreanu, 2022).

With this context in mind—and because most AI detectors are not free—this study seeks to monetarily quantify people's reliance on AI to detect AI. We introduced a **deepfake detection task** in the lab, where participants were asked to identify the proportion of AI-generated content in synthetic 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 with another human peer before revising their initial identifications. We have **4 research questions**, as follows,

1. *Are participants willing to pay more to use ChatGPT than to cooperate with human peers to detect deepfake news?*
2. *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 task (GPT-2 generated news)?*
3. *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 these assistance?*
4. *Does ChatGPT help people more effectively than Human peers in the tasks of deepfake detection?*

2. Experimental Design

2.1. Main Task

The flow of main task is shown in Figure 1. In the main task, participants were asked to finish 22 rounds of deepfake detection tasks, where the first half (Round 1 ~ Round 11) is named Part 1 and the second half (Round 12 ~ Round 22) Part 2. Before each part, participants submit their willingness to pay for access to a paid "External Assistance" (**CHAT** below) in the upcoming 11 rounds (then WTP1 for Part 1 and WTP2 for Part 2). Between the two parts, participants were shown the feedback of their performance in Part 1 .

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

$$AIpro = \frac{the\ length\ of\ AI\ generated\ part\ of\ the\ news}{the\ length\ of\ the\ news} \times 100, \quad (1)$$

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

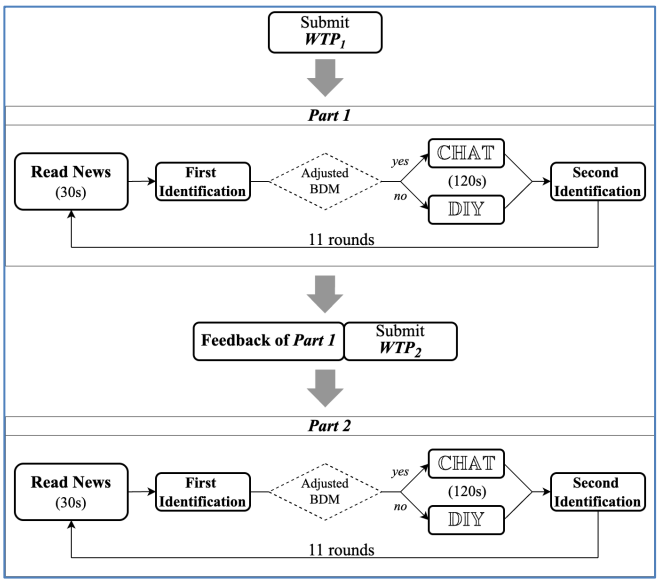

Figure 1: Main Task

For the deepfake news materials, we used two different Large Language Models (GPT2 & GPT4o) to extend some human-written news to 400-word length. 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" (**CHAT**) or on their own (**DIY**)—and report a final, revised identification (*2ndResp*).

Whether a participant can access **CHAT** or **DIY** in a given round was determined by an adjusted BDM procedure (Figure 2). Under the classical BDM (Becker et al., 1964), 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.

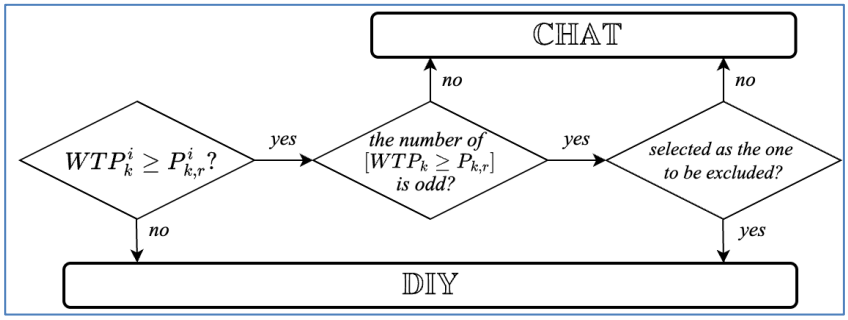

Figure 2: Adjusted BDM

For the paid "External Assistance", we offered participants two forms of **CHAT**:

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

Therefore, 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 follows,

- **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;

Each participant earned a fixed participation fee of 1,000 JPY plus a performance based bonus $\boldsymbol{\pi}$. 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 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 **CHAT** in round rn2, the price P_{rn2} was deducted.

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

, where the *inChat* = 1 if the participant accessed **CHAT** in round rn2 (0 otherwise).

2.2. Hypotheses

We assume none of the participants had experience with a deepfake detection task. Consequently, their first bid, WTP1, represents a prior valuation of the information they expect from CHAT before starting Part 1, whereas the second bid, WTP2, 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— and for ChatGPT in particular. 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.*

Perceived difficulty of tasks can influence participants' valuation of the outcome and their willingness to make a greater sacrifice to obtain it. 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 have

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

Participants may be unfamiliar with the task and the external assistance at first; thus, their actual usage experience could shape subsequent payment preferences. 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 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,

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

3. Main Results

The experiment was conducted in the laboratory at the Institute of Social and Economic Research (ISER) at the University of Osaka on January 23rd and 24th, 2025, and in the laboratory at the Research Institute for Socionetwork Strategies (RISS) at Kansai University on January 29th and 30th, 2025. We recruited 158 student participants, 37 out of whom were assigned to the AI2 treatment, 41 to the AI4, 37 to the HM2 and 43 to the HM4 treatment.

3.1. Performance

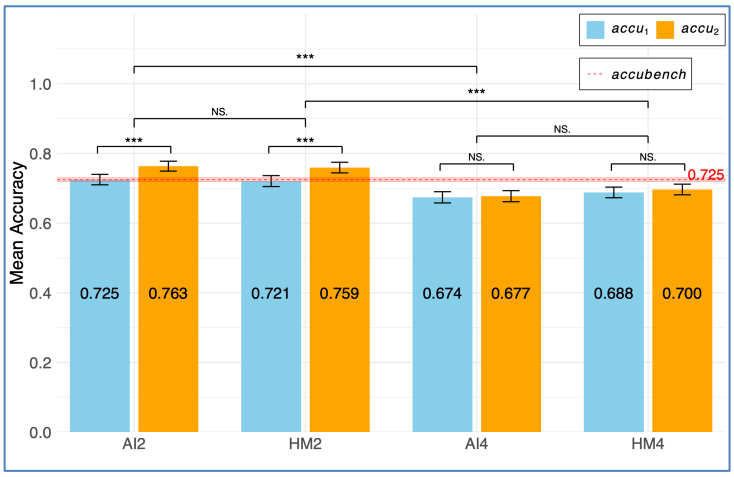

Figure 3: Performance Comparisons

Figure 3 reports comparisons of mean detection accuracy by treatment. Between-treatment comparisons use the Mann–Whitney U test (*$p < 0.05$, **$p < 0.01$, *** $p < 0.001$). On average, both the initial identification (accu1) and the final identification (accu2) are higher when detecting GPT-2 news than GPT-4o news.

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

We next examine participants' improvement from the initial to the final identification. To account for the ceiling (``threshold'') effect --- that is, the dependence of $accu_2$ on $accu_1$---we estimate OLS models by regressing the **proportional reduction in error** (PRE=$\frac{accu_2 - accu_1}{1 - accu_1}$). In all of the models, the coefficients on the two treatment indicators --- $inAI$ and $gpt4news$ --- are not statistically significant. By contrast, the coefficients on $inChat$ (= 1 if the participant accessed **CHAT** in one round and 0 otherwise) are negatively significant, while all of the coefficients of the interaction term are not significant. Therefore, we have

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

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

3.2. WTP

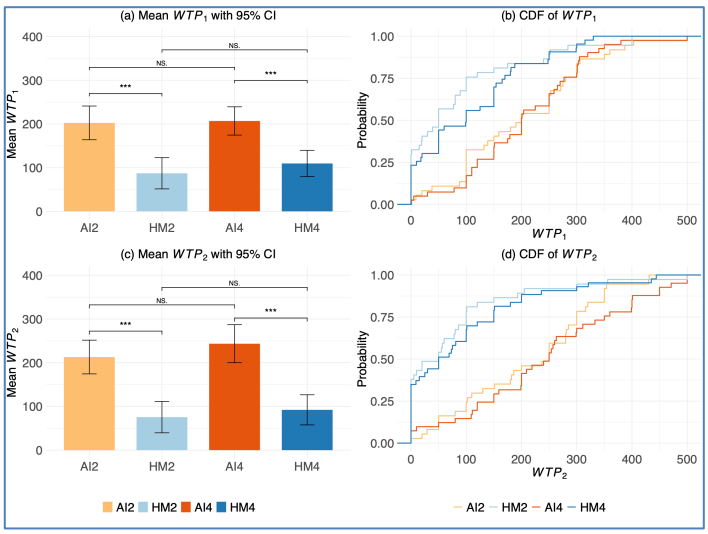

Figure 4: WTP Comparisons

Figure 4 displays WTP1 and WTP2 using bar charts and CDFs. The results of Kolmogorov – Smirnov test show that for both WTP1 and WTP2, 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. Therefore,

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.*

3.3. Feedback Effect

To examine feedback effects, we estimate Probit models of **the likelihood of raising WTP** (*WTPup*). 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, the task difficulty does not systematically affect *WTPup*. We also focus the extent to which participants increase their WTP for **CHAT** by regressing the **Relative Magnitude of WTP Adjustment** ($RltWTPup = \frac{WTP2-WTP1}{500-WTP1}$). OLS estimates show that AI treatment shows positive but generally insignificant coefficients, providing limited evidence that AI detector increases the WTP adjustments. Improvements given by DIY are negative and significant, indicating that participants who perform better on their own tend to lower their willingness to access **CHAT**.

Result 5: *When the external assistance is an AI detector, participants are more likely to increase their WTP for accessing it after experiencing the task.*

Result 6: *After experiencing the task, positive feedback from accessing **CHAT** increases the likelihood of raising WTP for accessing **CHAT**, whereas positive feedback from **DIY** decreases the likelihood of raising WTP for accessing **CHAT**.*

Reference

Becker, G. M., DeGroot, M. H. and Marschak, J. (1964), 'Measuring utility by a single response sequential method', Behavioral science 9(3), 226–232.

Passi, S. and Vorvoreanu, M. (2022), 'Overreliance on ai literature review', Microsoft Research 339, 340.