Do people rely on ChatGPT more than their peers to detect fake news? [*]

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

Abstract

In the era of rapidly advancing artificial intelligence (AI), understanding to what extent people rely on generative AI products (AI tools), such as ChatGPT, is crucial. This study experimentally investigates whether people rely more on AI tools than on their human peers in assessing the authenticity of misinformation. We quantify participants' degree of reliance using the weight of reference (*WOR*) and decompose it into two stages using the activation-integration model. Our results indicate that participants exhibit a higher reliance on ChatGPT than their peers, influenced significantly by the quality of the reference and their prior beliefs. In addition, we found that the reference source affects both the activation and integration stages, but the quality of reference only influences the second stage.

Keywords: ChatGPT, AI reliance, fake news identification, WOR, Heckman selection
JEL classification: C90; D83; D91

[*] This research has benefited from the financial support of (a) the Joint Usage/Research Center, the Institute of Social and Economic Research (ISER), and Osaka University, and (b) Grants-in-aid for Scientific Research Nos. 20H05631 and 23H00055 from the Japan Society for the Promotion of Science. The design of the experiment reported in this paper was approved by the IRB of ISER (#20231001) in October 2023, and the experiment is preregistered at aspredicted.org (#149838).

[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

Generative artificial intelligence (GAI), such as OpenAI's ChatGPT, has gained global attention. However, GAI also generates risks, particularly in the spread of misinformation like fake news. This study seeks to examine **whether people rely more on ChatGPT than on their human peers to detect fake news**. In our experiment, participants assessed the authenticity of misinformation and updated their judgments based on references from either ChatGPT or human peers. We compared participants' reliance across groups by the "weight of reference" (WOR) and applied the two-stage model of Vodrahalli et al. (2022) along with the Heckman correction (Heckman, 1974) to decompose reliance into activation and integration stages.

As a result, our findings indicate that participants exhibit a significantly higher degree of reliance on ChatGPT than on their peers. In the analysis of decomposing reliance, participants' assessment of the quality of reference did not affect the activation stage but the integration stage. Additionally, prior beliefs were found to strongly influence participants' reliance.

2. Experimental Design

2.1. Main Task

The main task consists of 30 rounds, with four stages per round, as shown in Figure 1.

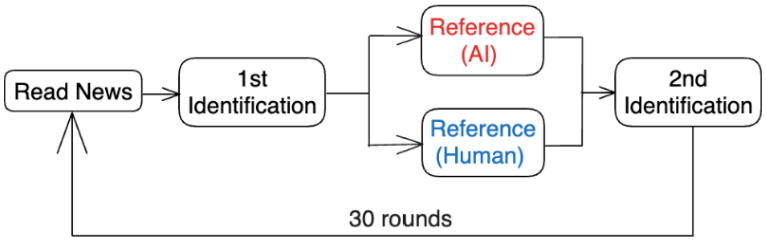

Figure 1: Main Task

In the **Read News Stage**, each participant read the news which were Japanese news collected from an open fake news dataset. The 30 pieces of news came in three types: totally real, totally fake, partially fake. The totally real news was written by humans, the totally fake news was generated by Google's GPT-2 Japanese model, and the partially fake news was the composition of real and fake. The proportion of the real part, $realr$,[1] for each piece of news, is defined as

$$realr^s = \frac{the\ length\ of\ real\ part\ of\ the\ news\ in\ round\ s}{the\ length\ of\ the\ news\ in\ round\ s}. \quad (1)$$

After reading the news, participants entered the **First Identification Stage**, where they used a slider to report a number between 0 and 100 for their initial identification ($response_1$). In the **Reference Stage**, participants were split into two groups: the AI group, which received a reference randomly selected from 24 ChatGPT responses, and the Human group, which received a randomly selected $response_1$ from another participant. In the final **Second Identification**

[1] Participants were informed that in this experiment, the "authenticity" they need to identify is defined as *realr*.

Stage, participants submitted their second identification ($response_2$).

Participants' final payoff consisted of a fixed fee (500 JPY) and an additional performance-based amount. The additional payoff was calculated from the accuracy of one randomly selected response out of 60 total responses (30 rounds × 2 responses). The additional payoff π was determined using the following equation, where R is the randomly selected response:

$$\pi = max\{0, 2300 - 0.3 \times (R - realr)^2\}. \quad (2)$$

2.2. Hypotheses

Studies on advice-taking used the weight of advice (WOA) to measure the degree to which people take advice (Önkal et al., 2009) in the form of a numerical estimate. In this study, we renamed it as "**weight of reliance**" (*WOR*), which is calculated by

$$WOR = \frac{response_2 - response_1}{ref - response_1}, \quad (3)$$

where *ref* denotes the references. This method provides a continuous outcome from 0 (completely ignoring the reference) to 1 (completely relying on it). We then hypothesize that participants in our experiment will rely more on ChatGPT than their peers:

H1: The *WOR* in the AI group is higher than that in the Human group.

As noted, the news materials used in the tasks included three types, with misinformation content ranging from 0% to 100%. We hypothesize that reliance on ChatGPT increases in more challenging tasks, such as assessing the authenticity of partially fake news versus totally fake news. This leads us to the second hypothesis:

H2: The *WOR* in the AI group increases for partially fake news compared to totally fake news.

3. Main Results

The experiment was programmed using Otree 5, conducted on November 7th and November 9th, 2023, in the laboratory at the Institute of Social and Economic Research (ISER) at Osaka University. We recruited 37 native Japanese-speaking students, with 17 assigned to the Human group and 20 to the AI group.[2] Each participant completed 30 rounds, resulting in a total sample size of 600 for the AI group and 510 for the Human group.[3]

3.1. Weight of Reliance

We initially used WOR as our primary analysis method , but when initial identification is close to the reference, WOR can exceed one or become infinite, distorting belief updates. While most studies cap the value, we excluded only infinite cases to preserve data integrity. This adjusted the sample size to 494 in the Human group and 562 in the AI group. Figure 2 shows the average

[2] A power analysis based on the result of a pilot experiment suggests that we need 17 participants in each group.
[3] In analyses, we corrected the standard error to account for multiple observations collected from the same participant.

WOR comparison between groups with 95% confidence intervals.

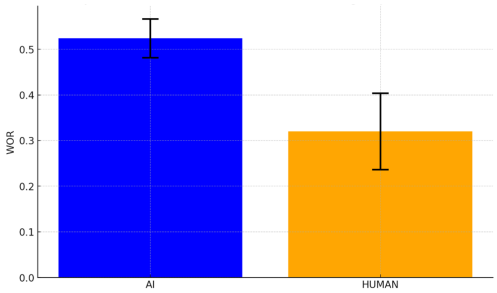

Figure 2: The Average WOR Comparisons

We regressed *WOR* on the treatment dummy, *inAI* (1 if the reference is from ChatGPT), along with control variables using an OLS model. Table 1 shows the results, where *aveRead* is the time spent reading each character, *timeidt1* and *timeidt2* represent the time spent on the first and second identifications, *roundnum* is the round number, and *accuref* denotes the quality (accuracy) of the reference, which defined as "1-normalized absolute error", that is, $1 - \frac{|Ref-real|}{100}$. The positive sign of *inAI* confirms our hypothesis **H1**, showing that participants relied more on ChatGPT than their peers. While time-related variables had no significant effect, the positive sign of *accuref* indicates a significant positive effect of reference quality on reliance.

Table 1：Source, Time, and Reference Quality

Var.	Estimate	S.E.	t value	Pr(>|t|)
inAI	**0.180**	**0.072**	**2.507**	**0.012***
aveRead	-0.535	0.338	-1.584	0.114
timeidt1	-0.006	0.008	-0.699	0.485
timeidt2	-0.001	0.002	-0.219	0.827
roundnum	-0.002	0.004	-0.523	0.601
accuref	**0.144**	**0.073**	**1.9981**	**0.048***

* p<0.05, ** p<0.01, *** p<0.001

3.2. News Type and Prior Beliefs

We represented news type in two ways: using *realr* (the proportion of real content) and two dummy variables, *isfake* and *isreal* (1 if the news is totally fake or real, respectively, and 0 otherwise). Combining these with *inAI* and *accuref*, we ran some OLS regressions on WOR. As a result, the type of news had no significant effect on reliance, leading to **the rejection of H2.**

In the survey, we assessed participants' prior beliefs by asking who they thought provided more accurate responses: "GAI," "Human," or "Not sure." We created a dummy variable, *priorcons*, set to 1 if participants in the AI (Human) group believed GAI (Human) was better, indicating consistency with the reference source. Using *WOR* as the dependent variable, OLS regressions on *priorcons* showed a positive effect, suggesting **participants relied more on references from sources they believed were better**, highlighting the impact of prior beliefs.

4. Decomposing Reliance

4.1. Processing Reference in Two Stages

As noted, WOR samples need adjustment when the denominator of equation (3) approaches zero, as this may not accurately reflect belief updates. To address this, we applied the activation-integration model (Vodrahalli et al., 2022) with Heckman correction (Heckman,1974) to decompose reliance into two stages: the first (activation) stage, where participants decide whether to use the advice, and the second (integration) stage, where they determine the extent of its use. This method allows us to include the entire sample for more accurate results.

4.2. Activation Stage

In our analysis, all observations in the activation stage are defined as follows,

$$Acti = \begin{cases} 1 & if\ response_1 \neq response_2 \\ 0 & if\ response_1 = response_2 \end{cases}, \tag{4}$$

specifically, a participant is considered **activated** if they altered their initial response after receiving a reference and **not activated** if they maintained their initial response. To further analyze, we ran three Probit regressions on *Acti*. In addition to the source (*inAI*) and reference quality (*accuref*), we examined the effects of time, news type, and prior beliefs, and *diffref*.[4] Based on the results, we identified **three key factors** influencing activation: **the reference source** (*inAI*), **prior beliefs** (*priorcons*), and **the gap between $\boldsymbol{ref}$ and $\boldsymbol{response_1}$** (*diffref*).

4.3. Integration Stage

In the second stage, our focus shifts to participants who are activated. To quantify the extent of reference utilization, we constructed a continuous variable as follows,

$$consref = \begin{cases} response_2 - ref & if\ ref > response_1 \\ |response_2 - ref| & if\ ref = response_1, \\ ref - response_2 & if\ ref < response_1 \end{cases} \tag{5}$$

which describes **the consistency with the reference**. Therefore, *consref* > 0 indicates that a participant moved their second identification ($response_2$) on the slider beyond the reference point (*ref*), **overutilizing** the reference. *consref* = 0 means they matched $response_2$ exactly to the reference on the slider, **totally utilizing** it. *consref* < 0 suggests they moved $response_2$ less than needed to reach the reference, indicating **underutilization**. Here, we further applied the following Heckman selection (Heckman, 1974) model:

Activation **(Selection):**

$$Acti = \alpha_0 + \alpha_1 \cdot inAI + \alpha_2 \cdot diffref + \alpha_3 \cdot priorcons + \varepsilon, \tag{6}$$

Integration **(Outcome):**

$$consref = \beta \cdot X + \gamma \cdot imr + u, \tag{7}$$

where equation of (6) is a probit regression model, in which we choose *inAI, diffref*, and *priorcons*

[4] *diffref* represents the distance between the reference and the initial response, defined as $|ref - response_1|$

as independent variables as they have been proven to affect *Acti* significantly. Equation of (7) is an OLS model, and X consists of *inAI*, *priorcons*, and other variables. *imr* denotes the **inverse mills ratio**, calculated by $imr = \frac{pdf(\widehat{Acti})}{cdf(\widehat{Acti})}$, and it can be concluded that the selection effect exists if the coefficient of *imr* is significant. We applied this sample selection model to our analysis as we considered that there may exist a reservation level of *inAI*, *diffref*, and *priorcons*. If these variables do not reach a certain threshold, a participant might not be activated.

Table 2: Heckman MLE & Heckit

	OLS	MLE		Heckit	
	Integration	Activation	Integration	Activation	Integration
inAI	**6.013***** (1.77)	**0.532**** (0.17)	**10.519***** (2.53)	**0.579***** (0.22)	**22.861***** (1.91)
priorcons	2.939 (1.61)	**0.391*** (0.16)	**6.394*** (2.51)	**0.497*** (0.22)	**16.896***** (2.19)
accuref	**16.854***** (3.12)		**11.074***** (2.38)		**5.940*** (2.59)
...	...	...	...	...	...
diffref		**0.049***** (0.00)		**0.019***** (0.00)	
imr			$[\boldsymbol{\rho \cdot \sigma_\varepsilon}]$***		**73.476***** (7.28)
$arctanh(\rho)$			**2.263***** (0.20)		
$ln\ (\sigma_\varepsilon)$			**2.858***** (0.07)		

* p<0.05, ** p<0.01, *** p<0.001; $\text{arctanh}(\rho) = \frac{1}{2}ln[1+\rho]$, where ρ is the correlation coefficient between *Acti* and *consref*; σ_ε is the standard error of the residual in the activation equation. $imr = \rho \cdot \sigma_\varepsilon$;

We used both the methods of maximum likelihood estimation (MLE) and two-step estimation (Heckit). The results are shown in the third and fifth columns of Table 2. First, the significance of the coefficients for *imr* confirms the presence of selection bias. Second, the reference source influences participants' use of references in the integration stage, with ChatGPT prompting greater reliance. Third, prior beliefs affect both stages—participants rely more on references from sources they perceive as better for the task. Lastly, the positive effect of *accuref* indicates that participants weigh reference quality heavily, even though it doesn't affect their activation decision.

Reference

Heckman, J., 1974. Shadow prices, market wages, and labor supply. Econometrica: Journal of the Econometric Society, 679-694.

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

Vodrahalli, K., Daneshjou, R., Gerstenberg, T. and Zou, J., 2022, July. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (763-777).