生成 AI 利用によるフェイクへの反応の変化：
フェイクニュースのオンライン実験

石原　卓典[a]

要約

本研究では，日本国内の成人モニター1,169 名を対象としたオンライン実験により，生成 AI の提示および利用がフェイクニュース判断に与える影響を検証した．調査では，参加者を統制群・生成 AI 情報群・生成 AI 利用群の 3 群にランダムに割り当て，10 問の正誤問題に回答させた．その結果，統制群に比べて両介入群の正答率についてはいずれも有意に増加する傾向がみられた．一方で，出力に対する信頼度については生成 AI 利用群の方が有意に高く，生成 AI を「自ら利用した」経験が提示のみよりも出力への信頼感を高めることが明らかになった．さらに，Sentence-BERT を用いた意味的類似度分析の結果，生成 AI 出力とファクトチェック記事との類似度が高い設問ほど介入効果が大きいことが確認された．本研究は，生成 AI がフェイクニュース判断を改善する可能性を示すと同時に，利用経験がユーザーの主観的信頼を高めることを実証的に示した．

JEL 分類番号：C91，D83，O33
キーワード：フェイクニュース，生成 AI，オンライン実験，意味的類似度

[a] 京都先端科学大学国際学術研究院　ishihara.takunori@kuas.ac.jp

本研究を実施するにあたり，京都先端科学大学の「2024 年度先端研究」による研究助成を受けている．また、本研究の遂行および論文執筆にあたり、著者には利益相反に該当する事項はない。

1. イントロダクション

近年，フェイクニュースが社会に与える影響は深刻化している．一例として，選挙期間中における SNS 空間でのフェイクニュースの拡散により，投票行動に影響が出る可能性が懸念されている．Kartal and Tyran (2022) は，偽情報と有権者の過信が民主的選択の質を損なうことを示している．また，このようなフェイクニュースの受容については，Thaler (2024) では信念に基づく推論がフェイクニュースの受容に影響することを実験的に明らかにしている．こうした研究は，フェイクニュースが単なる情報問題にとどまらず，政治や経済，制度などの根幹に影響することを示している．

フェイクニュースに対処する手段として，ナッジを用いたアプローチが注目を集めている．Pennycook et al. (2021) は，ニュース共有時に正確さを意識させる介入が偽情報の拡散を抑制することを示した．さらに，Chan et al. (2025) は多国比較の実験を通じて accuracy nudge の効果が文化や制度的文脈に依存することを明らかにしている．これらの研究は，ナッジが情報判断の補正手段として有効であることを強く示唆している．

一方，近年急速に普及する生成 AI は，フェイクニュースの生成を容易にするリスクを抱えると同時に，誤情報の検出や評価に資する可能性も持つ可能性がある．近年は「AI 検索」という言葉が登場し，検索や意思決定の過程で生成 AI を利用する新しい行動様式が広がりつつある．この変化は，人々の情報処理における AI の役割が中心化していくことを示している．

生成 AI にはハルシネーションの問題があるものの，うまく利用することで私たちの認知バイアスの削減に貢献する可能性が考えられる．生成 AI は部分的にではあるものの，人間が系統的に間違う問題に対して人間と同等かそれ以上の確度で問題に正解するということが指摘されている（Chen et al. 2023）．また，生成 AI は対話的に利用することが可能であるため，ナッジよりもより利用者ごとに個別化されているため，利用者にその情報を受け入れてもらえる可能性が高くナッジと比較してより大きな効果を持つ可能性がある．

本研究では，オンライン実験を行い，フェイクを含む正誤問題を設定し，異なる条件の下で回答を得ることで，生成 AI の利用が情報判断に与える影響を検証する．

2. データ

2025 年 2 月にオンライン調査会社の登録モニターを対象にアンケート調査およびオンライン実験を実施した[2]. 調査対象は，日本国内に居住する 20 歳－69 歳までの男女個人とし，性別が半数ずつ，年代が均等になるようにサンプリングを行った．また，本調査では生成 AI

[2] 本研究のオンライン実験を実施するにあたり，京都先端科学大学倫理審査委員会による倫理審査を受け，承認を得ている（審査番号：24EB02）.

（ChatGPT）を利用して回答を行う項目を含んでいるため，調査時の混乱を避けるために生成 AI を全く利用したことのない回答者は除外した．その結果，合計 1,200 名から回答を得た．以下の分析では，回答の不備や不適切な回答（同一選択肢を繰り返す等）が確認された 31 名を除外し，最終的に 1,169 名のデータを用いている．

アンケート調査では，以下の 3 種類の設問を設定した．第一に回答者の生成 AI 利用状況を把握する質問（6 問），第二にフェイクニュースを含む正誤問題（10 問），第三に，回答者の社会経済的属性や心理的属性に関する質問（17 問）である．フェイクニュースを含む正誤問題については，日本ファクトチェックセンターが公開するファクトチェック記事に基づいて作成し，各設問は「正しい／わからない／誤っている」の三択形式として，設問順序をランダムに提示した．アンケート回答に対して一律 100 円の報酬を付与したうえで，正誤問題に正答するごとに追加で 10 円の追加報酬を提供した（最大 100 円）．

正誤問題では，1 つの統制群（C）と 2 つの介入群のうち，いずれかの群にランダムに割り振り，割り振られた群の条件の下で回答を求めた．介入群として，生成 AI 情報群（T1）と生成 AI 利用群（T2）を設けた．生成 AI 情報群では，実験者が事前に ChatGPT を用いて生成した約 400 字の解説文を提示したうえで回答を求めた．生成 AI 利用群では，回答者自身が ChatGPT を利用して正誤問題に解答することを求め，その際に用いた質問文と生成 AI の出力結果を併せて提出してもらった．なお，生成 AI 情報群および生成 AI 利用群に割り当てられた回答者には，正誤問題終了後に ChatGPT の出力内容に対する信頼度を 0～100%の数値で回答してもらった．

最後に，調査の結果得られた社会経済的属性変数を用いて群間比較を行い，それらの変数の観点から適切にバランシングされているかを検討した．バランスチェックの結果，一部の項目については群間で有意な差がみられたが[3]，おおむね適切にバランスが取れていることを確認した．

3. 分析

3.1. 平均介入効果

本節では，まずロジット・モデルを用いて限界効果を推定した．被説明変数は，個人 i が設問 j に正答したか否かの二値変数である（$y_{ij}=1$ であれば正答，$y_{ij}=0$ であればそれ

[3] 生成 AI の有料版の利用状況（p 値=0），「生成 AI の出力結果は人間の考えや意見に近いと思う」（p 値=0.031），CRT（認知反射テスト）の正答率（p 値=0）で 3 群間の有意差がみられている．これらの属性変数については，のちの分析では回帰式に含め，群間のずれを補正している．

以外を表す)．設問は全部で 10 問あり，同一の個人が複数設問に回答しているため，各設問の回答をプールしたデータを用いて推定を行った．

分析で用いた推定モデルは以下のように定式化することができる．

$$U_{ij} = \tau_k D_{ik} + \beta X_{ij} + \gamma_j + \epsilon_{ij}$$

ここで，D_{ik} は個人 i が介入群 k に割り当てられていれば 1，そうでなければ 0 をとる介入ダミーである．また，X_{ij} は個人属性を含む共変量ベクトルを示している．γ_j は設問固定効果を表し，ϵ_{ij} は誤差項を表している．また，τ_k は介入 k による平均介入効果を示し，βは共変量の係数ベクトルを示している．

誤差項がロジスティック分布に従うと仮定すると，選択確率は

$$\Pr(y_{ij} = 1 | D_{ik}, X_{ij}, \gamma_j) = \frac{\exp(\tau_k D_{ik} + \beta X_{ij} + \gamma_j)}{1 + \exp(\tau_k D_{ik} + \beta X_{ij} + \gamma_j)}$$

と表すことができる．

＜図 1：介入の限界効果＞

各介入による限界効果を推定した結果を図 1 に示している．まず，統制群における平均正答率は約 61%であった．そのうえで，生成 AI 情報群（T1）では正答確率が統制群に比べ約 11.8%高く，生成 AI 利用群（T2）では約 11%高かった．生成 AI 情報群と生成 AI 利用群の間では限界効果に有意な差は見られなかった（p 値=0.6906）．

3.2. 意味的類似度の分析

さらに，本研究では生成 AI 出力の「質」が介入による限界効果にどのように寄与しているかを分析するため，Sentence-BERT を用いた意味的類似度スコア（BERT スコア）を算出し，それを用いて追加分析を行った．BERT スコアを計算するにあたって，各設問の基としたファクトチェック記事を真値テキストとし，生成 AI による出力結果を予測テキストとした．これらのテキストのペアをそれぞれ Sentence-BERT によりベクトルに変換し，コサイン類似度を計算した．コサイン類似度は値が 1 に近いほど 2 つのテキストが意味的に類似していることを示し，逆に-1に近いほど意味的に異なることを表している．以下では 3.1 節で行ったロジット分析に BERT スコアと介入ダミーの交差項を含めて推定を行った[4]．

＜図 2：介入と類似度についての限界効果＞

[4] 統制群については生成 AI による出力が存在しないため，分析では BERT スコアが 0 であると仮定した．

図 2 では，各介入ダミーと BERT スコアの交差項を含めたうえでの限界効果を示している[5]．推定結果より，BERT スコアが増加するに従い，どちらの介入についても正答率が増加することがわかる．しかし，介入の違いによる差は見られなかった（p 値=0.9599）．

3.3. 信頼度の分析

最後に，参加者がどの程度生成 AI による出力を信頼したかについて分析する．今回の調査では，生成 AI 情報群と生成 AI 利用群に対して，正誤問題への回答の後にそれまで出力された生成 AI の回答に対する信頼度を一括して 0 から 100 の間の数値で回答してもらった．

分析の結果，生成 AI 情報群での平均信頼度は 50.6%であり，生成 AI 利用群では生成 AI 情報群と比べ，8.2%信頼度が高くなる傾向がみられた（p 値=0.013）．正誤問題に対する正答率についてはどちらの介入についても同等の効果がみられるが，自分で生成 AI を利用する生成 AI 利用群では，単に情報を提示される生成 AI 情報群に比べて，出力への信頼度は高くなる傾向がみられることが分かった．

4. 結論

本研究は，生成 AI の提示および利用が，フェイクニュース判断に対して有意な効果を持つことを明らかにした．具体的には，統制群に比べて生成 AI 情報群・生成 AI 利用群の正答率が上昇し，特に利用群で効果が大きかった．さらに，Sentence-BERT を用いた意味的類似度分析の結果，AI 出力のテキストとファクトチェック記事の類似度が高い設問ほど，介入効果が大きいことが確認された．加えて，AI 出力に対する信頼度を回答者に尋ねた結果，生成 AI 利用群では生成 AI 情報群と比べて平均的に高い水準の信頼が示された．実際の正答率では両群で差がみられないものの，生成 AI を「自ら利用した」経験が提示のみよりも出力への信頼感を高めることが明らかになった．

先行研究

[1]. Chan, M., Yi, J., Vaccari, C., & Yamamoto, M. (2025). A cross-national examination of the effects of accuracy nudges and content veracity labels on belief in and sharing of misleading news. *Journal of Computer-Mediated Communication*, *30*(4), zmaf009.

[2]. Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., & Jenkin, T. (2025). A

[5] BERT スコアの平均値は生成 AI 情報群で 0.71（s.d.=0.10），生成 AI 利用群では 0.67（s.d.=0.15）であった．

manager and an AI walk into a bar: does ChatGPT make biased decisions like we do?. *Manufacturing & Service Operations Management*, *27*(2), 354-368.

[3]. Kartal, M., & Tyran, J. R. (2022). Fake news, voter overconfidence, and the quality of democratic choice. *American Economic Review*, 112(10), 3367-3397.

[4]. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590-595.

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

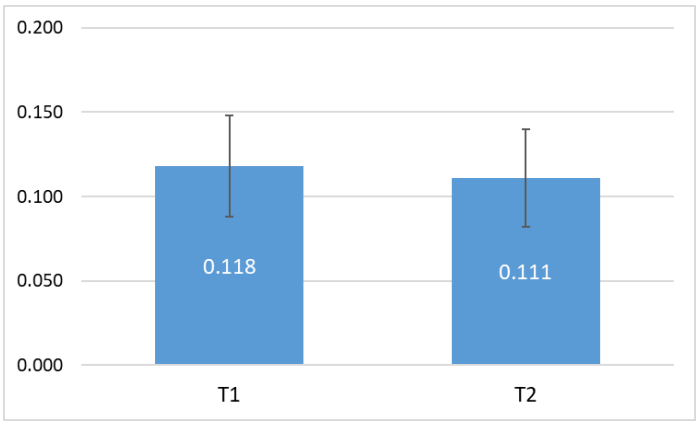

図 1：介入による限界効果

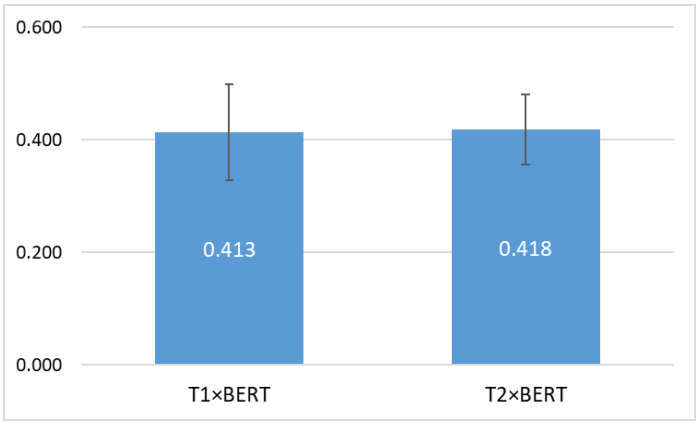

図 2：介入と類似度についての限界効果