危険選好と参照点依存をめぐる人間と AI の比較研究：ペルソナ分析とチューニング*

岩本涼太[a]　石原卓典[b]　依田高典[c]

要約

本研究では，人間と生成 AI における危険選好および参照点依存の違いについて，実証的に検討を行った．全国 4,838 名を対象とした Web 調査を実施し，それと同一条件で属性情報に基づくペルソナを用いた生成 AI の応答を取得した．分析の結果，利得では人間と AI の選好傾向は類似しており，いずれも危険回避的な選択を示した．他方，損失では AI において危険愛好的な傾向が顕著に現れ，人間よりも性別や年齢・所得といった属性による影響を強く受けていた．さらに，人間の選択データに基づくファインチューニングを通じて AI を再訓練することで，学習後の AI の応答傾向が人間の選好分布により近づき，とりわけ損失においてその類似性が高まった．ワッサースタイン距離を用いた分布比較からも，ファインチューニングが AI の人間への近接性を高めることがわかった．

JEL 分類番号：D91, C91

キーワード：危険選好，参照点依存，生成 AI，ファインチューニング，ワッサースタイン距離

---

* 本論文に関して，開示すべき利益相反関連事項はない．

[a] 京都大学公共政策大学院 iwamoto.ryota.82a@st.kyoto-u.ac.jp

[b] 京都先端科学大学国際学術院 ishihara.takunori@kuas.ac.jp

[c] 京都大学大学院経済学研究科 ida@econ.kyoto-u.ac.jp

1. イントロダクション

ヒューマン・イン・ザ・ループ（Human-in-the-Loop, HITL）は，経済的意思決定において人間と人工知能（AI）を組み合わせる枠組みとして注目を集めている（Rahwan 2018, Rahwan et al. 2019）．特に，AI と人間におけるバイアス構造の比較は，HITL の意義を理解するうえで重要な研究領域である．異なるバイアス構造を理解し，AI と人間の間で相互に補正し合う設計を導入することは，望ましい意思決定を実現するうえで不可欠である．そのため，AI と人間の判断傾向の類似点および相違点に着目した比較研究が近年急速に増加している．

ファインチューニングとは，既存の基盤モデルに対して特定のタスクや領域に関する追加データを用いて再訓練を行う手法であり，モデルの出力傾向や性能の目的に応じた調整を行うことが可能である．ファインチューニングの前後において，生成 AI が人間の意思決定に内在するバイアスをどの程度再現することができるかを明らかにすることも，行動経済学および AI 応用の双方にとって重要な研究課題である．

本研究は，危険選好の感応度逓減と損失回避性と関連が深い参照点依存に着目し，人間と生成 AI（GPT-4o）の意思決定に伴うバイアスの実証的比較検討を目的とする．具体的には，日本国内で Web 調査（N=4838）を実施し，個人属性（性別，年齢，所得）と回答傾向との関連性を分析する．そして，得られた属性情報より仮想的なペルソナを設計し，異なる temperature 条件の下で GPT-4o に同一の設問を提示する．そのうえで，AI の応答と人間の選好傾向を定量的に比較し，両者の意思決定構造の相違点および一致度を検証する．さらに，人間の選好に基づくファインチューニングを通じて AI を再訓練し，ファインチューニング前後での AI と人間の選好分布の類似度を，分布の形状の差異把握に特化したワッサースタイン距離を用いて分析する．

生成 AI と人間の意思決定や認知バイアスの比較に関する先行研究は，生成 AI が特定の条件下において人間の行動や判断を再現し得ることを示す一方で，その再現性が属性情報や文脈，設問形式，さらにはプロンプトやペルソナの設計といったモデル設定に大きく依存することを明らかにしている（Park et al. 2024, Ross et al. 2024, Chen et al. 2025）．

他方で，これらの先行研究は，生成 AI と人間の意思決定を比較する際に，再現性の評価指標を十分に精緻化しているとは言いがたい．また，平均的な傾向の再現性は多くの研究で検証されているものの，人間の属性で条件づけた評価を行っている研究は限定的である．たとえ平均的には人間と類似した傾向を示していたとしても，より細分化された条件下，あるいは個人レベルに近い評価においては，異なる応答傾向が顕在化する可能性がある．さらに，生成 AI に人間の回答をフィードバックすることにより，再現性の改善が見込まれるものの，これを実装した研究は非常に限られている．以上を踏まえると，生成 AI と人間行動の類似性・相違性をより厳密に評価する上で，本稿が提示する評価方法は，既存研究と比較して分析手法の妥当性および再現性の精度の両面において優位性を持つものと考えられる．

2. 調査設計

2.1. Web 調査および生成 AI への調査の設計

我々は 2024 年 12 月に，インターネット調査会社を通じて，20 歳から 65 歳までの国内在住者（N=4838）を対象に Web 調査を実施し，個人属性，認知バイアスや心理特性に関する応答を取得した．本稿では，分析の焦点を絞るために，個人属性については性別と年齢・所得に，心理特性については危険選好および参照点依存のみに着目して分析を行う．提示した設問は，Tversky and Kahneman（1988） の問題設定に依拠した以下の2質問である．

質問1．あなたは現在の富に上乗せして 30,000 円もらったうえで，以下のどちらかの選択肢を選ぶように言われました．あなたはどちらの選択肢を選びますか？

選択肢1．確実に 10,000 円もらえる

選択肢2．50％の確率で 20,000 円もらえて，50％の確率で何ももらえない

質問2．あなたは現在の富に上乗せして 50,000 円もらったうえで，以下のどちらかの選択肢を選ぶように言われました．あなたはどちらの選択肢を選びますか？

選択肢1．確実に 10,000 円失う

選択肢2．50％の確率で 20,000 円失い，50％の確率で何も失わない

生成 AI に対しては，以下の方法で認知バイアスに関する調査を行った．まず， Web 調査で収集した人間の性別・年齢・所得の3つの属性データをプロンプト形式で GPT-4o に入力し，特定属性を有する仮想的な人格（ペルソナ）を生成 AI に付与し，その上で，Web 調査と同一の選択課題を提示した．さらに，生成される応答のばらつきを制御するため，生成テキストのランダム性を調整するパラメータである "temperature" を3段階に設定し，それぞれの条件下で応答を取得した．

2.2. バイアスの推定手法

本分析では，二値の回答を被説明変数，性別・年齢・世帯年収の3つの属性を説明変数とし，最尤法による推定を行う．ここで，各選択肢の効用に付随する誤差項が第一種極値分布（Type I Extreme Value Distribution）に従うと仮定すると，誤差項の差はロジスティック分布に従うため，ロジット・モデルによる推定が適切である．ロジット・モデルの選択確率は（1）式で与えられる．

$$P(Y_i = 1 | X_{1i}, X_{2i}, X_{3i}) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3 X_{3i})}} \ . \qquad (1)$$

ここで，iは各回答者を表している．Y_iは回答者iの選択肢を表すダミー変数であり，選択肢1を選択した場合 1 を，選択肢2を選択した場合 0 をとる．X_{1i}は回答者iの性別を表すダミー変数であり，

個人iが女性であれば 1 を，男性であれば 0 をとる．X_{2i}は個人iの年齢（10 歳単位）である．X_{3i}は個人iの世帯年収（100 万円単位）である．

3. 人間と学習前 AI のバイアス比較

第一に，選択傾向の概要について説明する．まず，人間の Web 調査の結果を説明する．利得の質問1では，選択肢1（確実に 1 万円を得る）を選択した割合は 88.4％であり，利得における危険回避的傾向を示唆している．他方，損失の質問2では，選択肢1（確実に 1 万円を失う）を選択した割合が 57.7％であり，一定程度危険愛好的な選択も見られたが，明確な傾向とは言いがたい．次に，デフォルト値の temperature＝1.0 における生成 AI の応答結果を示す．質問1では，確実な利得を選択した割合が 91.1％に達し，危険回避的な傾向が明確に観察された．他方，質問2では，確実な損失を選んだ割合は 11.0％にとどまり，多くの応答が確率的な損失を選択しており，損失における危険愛好性が示された．Web 調査の結果と生成 AI（temperature＝1.0）の結果を比較すると，質問1では両者とも 9 割近くが選択肢1を選んでおり，利得における危険回避性が共通して確認される．他方，質問2の選択率の違いは，損失における選好が人間と AI で異なる可能性を示唆している．さらに，両質問での選択肢1の選択率を比較すると，生成 AI では損失においてより強い参照点依存がうかがえるのに対し，人間の Web 調査では相対的に弱い参照点依存がみられる．

第二に，限界効果について述べる．まず，人間の Web 調査の推定結果を概観する．利得の質問1では，選択肢1を選ぶ確率が，女性である場合に 4.1％増加し，年齢が 10 歳高くなることで 1.5％増加，所得が 100 万円増加することで 0.5％低下するという傾向が見られた．他方，損失の質問2では，選択確率がそれぞれ年齢については 2.1％増加し，所得は 0.8％低下，性別の影響はないという結果が得られた．次に，生成 AI（temperature＝1.0）の応答結果について検討する．質問1では，選択肢1を選ぶ確率が，性別については 3.3%増加し，年齢については 5.7%増加，所得については 2.7%低下する傾向が示された．質問2では，選択確率がそれぞれ性別については 2.8％増加し，年齢については 2.3％増加，所得については 0.6％低下する傾向が確認された．人間の Web 調査の結果と生成 AI（temperature＝1.0）の結果を，デルタ法を用いて比較すると，両者の限界効果には全体として統計的に有意な差があるとは言いがたいことが示された．そして，temperature の違いによる生成 AI の応答結果の変化を検討すると，いずれの temperature 設定においても，両質問および全属性に対して，有意水準 1％で統計的に有意な限界効果が確認された．また，temperature の違いによる限界効果の大きな変化は確認されなかった．

4. 生成 AI に対する学習と学習後 AI のバイアス分析

4.1. ファインチューニングの解説

本研究の第二の目的は，現実の人間の回答結果を生成 AI に学習させ，人間のバイアス傾向を

生成 AI で再現すること，そして学習前後の AI と人間の近接距離を測ることにある．我々は生成 AI に対する学習手法として，大規模データによる事前学習済みモデルを別のデータセットを用いて特定のタスクやドメインに特化させるための再訓練手法である，ファインチューニングを用いた．

我々は訓練用データとして，ロジット・モデルにより推定した各回答者の予測確率を用いた．これは推定されたロジット・モデルを用いて各サンプルについて選択肢1を選ぶ確率$\hat{P}(Y_i = 1|\boldsymbol{X})$を算出したものである．そして学習後 AI に対しても，学習前 AI と同様に，Web 調査と同一の属性（性別・年齢・所得）を持つ 4,838 件分のペルソナを生成 AI 上に再現し，2つの質問に対する応答を取得した．さらに，得られた応答に，2.2 節で述べたロジット・モデルを適用し，危険選好および参照点依存の傾向を分析した．

4.2. 人間と学習前 AI・学習後 AI のバイアス比較

まず，選択傾向の概要について説明する．学習後 AI の結果を確認すると，利得の質問1では，選択肢1を選択した割合は 96.1％であり，人間，学習前 AI と比べても高い選択率がみられた．これは利得における高い危険回避的傾向および危険回避性の過学習を示唆している．他方，損失の質問2では，選択肢1を選択した割合は 40.4％であり，損失における危険愛好性は強い傾向とは言えないが，学習によって人間の選択率に近づいており，損失における危険愛好性について適切に学習がなされたといえる．

次に，限界効果について述べる．学習後 AI の結果を確認すると，利得の質問1では，選択肢1を選ぶ確率は，性別については 1.9％増加，年齢については 3.0％増加，所得については 1.4％低下する傾向が見られた．他方，損失の質問2では，選択確率が性別については 26.4％増加し，年齢については 13.3％増加するという結果が得られた． 特に質問2においては，性別および年齢について大きな限界効果が確認された．このことから，生成 AI に学習させることにより，性別および年齢が損失の危険愛好性に与える影響が大きくなることが分かる．

最後に，学習前後における生成 AI と人間の意思決定傾向の近接度合について報告する．近接度の定量的な評価にあたり，確率分布間の距離を測定する指標として，2 次ワッサースタイン距離（Wasserstein distance of 2 order）を採用した．2 次ワッサースタイン距離は，最適輸送理論に基づき，2つの確率分布間における質量を移動させるための最小コストに着目して定義される．この指標により，学習前後の生成 AI と人間の選択の分布の差異を，選択確率分布間の距離として定量的に把握することが可能となる．

表1には，3者の選択確率の分布間における 2 次ワッサースタイン距離を示している．これより，学習を通じて生成 AI は人間の意思決定傾向に近づいており，特に損失においてその改善度が大きいことが明らかとなった．これは学習プロセスが生成AIのバイアス構造に影響を与え，より人間らしい判断特性を形成するうえで有効であったことを示唆している．

表 1　三者間の二次ワッサースタイン距離（AI は全て temperature=1.0 の値）

	人間-学習前 AI	人間-学習後 AI	学習前 AI-学習後 AI
利得（N=4,838）	0.135	0.123	0.081
損失（N=4,838）	0.467	0.263	0.353
合算（N=9,676）	0.343	0.198	0.253

5. 考察・結論

本研究では，危険選好および参照点依存に関する人間と生成 AI（GPT-4o）の意思決定傾向を比較し，特に属性情報によるペルソナ設定が AI の応答に与える影響を検討した．そして，人間の意思決定傾向を生成 AI に学習・再現させることを目指し，学習による人間への接近性を測定した．

人間と学習前 AI を比較すると，利得においては同様に危険回避的な傾向を示した．他方，損失では，生成 AI が人間よりも顕著に危険愛好的な選好を示すことが明らかとなった．加えて，個人属性の影響を考えると，生成 AI の方が選好のバイアスを強調しやすい性質があると考えられる．

人間と学習後 AI を比較すると，学習により生成 AI が人間の弱い損失回避傾向に接近することが明らかとなった．特に損失においては両者間の距離を大幅に縮めることに成功し，全体的にも学習が概ね適切に行われたことが示された．他方，利得においては危険回避性の過学習が，損失においては個人属性の影響の強調がそれぞれ見られ，学習に際して改善の余地があるといえる．

参考文献

[1]. Chen, Y., S. N. Kirshner, A. Ovchinnikov, M. Ovchinnikov, and T. Jenkin, 2025. A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do? Manufacturing & Service Operations Management 27, 2, 354-368.

[2]. Park, J.S., C.Q. Zou, A. Shaw, B.M. Hill, C. Cai, M.R. Morris, R. Willer, P. Liang, and M.S. Bernstein, 2024. Generative agent simulations of 1,000 people. arXiv preprint, arXiv:2411.10109.

[3]. Rahwan, I., 2018. Society-in-the-Loop: Programming the Algorithmic Social Contract. Ethics and Information Technology 20, 5–14.

[4]. Rahwan, I., M. Cebrian, N. Obradovich et al., 2019. Machine behaviour. Nature 568, 477–486.

[5]. Ross, J., Y. Kim, and A.W. Lo, 2024. LLM economicus? mapping the behavioral biases of LLMs via utility theory. arXiv preprint, arXiv:2408.02784.

[6]. Tversky, A. and D. Kahneman, 1988. Rational Choice and the Framing of Decisions. D. E. Bell, H. Raiffa, and A. Tversky ed., In Decision Making: Descriptive, Normative, and Prescriptive Interactions. Cambridge University Press, Cambridge, US.