



京都大学大学院経済学研究科
ディスカッションペーパーシリーズ

損失回避性をめぐる人間とAIの比較 研究：ペルソナ分析とチューニング

岩本涼太 京都大学 公共政策大学院 学生
石原卓典 京都先端科学大学 経済学部 准教授
依田高典 京都大学大学院 経済学研究科 教授

No. J-25-002

2025 年 4 月

〒606-8501
京都市左京区吉田本町
京都大学大学院経済学研究科

損失回避性をめぐる人間と AI の比較研究：

ペルソナ分析とチューニング

2025 年 4 月 10 日

中間報告版

本稿は、チューニング分析が未了の段階における中間報告である。今後、チューニング分析の完了を経て内容を改訂し、日本語版および英語版の最終報告版として公表する予定である。

岩本涼太* 石原卓典† 依田高典‡

要旨

本研究は、危険選好および損失回避性に関する人間と生成 AI (GPT-4o) の意思決定傾向を比較するものである。日本国内の成人 4,838 名に対する Web 調査を実施し、性別・年齢・年収といった属性が選好に与える影響を分析したうえで、同一の属性情報をもとにペルソナを構成し、生成 AI に同様の選択課題を提示した。その結果、利得局面では人間と AI の選好は類似していたが、損失局面では AI が人間よりも強い損失回避性を示した。また、生成 AI は属性の影響を人間よりも強調する傾向を持ち、temperature 設定によって応答の一貫性や現実性が変化することも確認された。

JEL 分類番号: D91, C91

キーワード: バイアス, 損失回避性, 危険選好, 生成 AI, ペルソナ

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

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

‡ 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。近年、AI による判断の自動化が進展する中で、その正当性や説明可能性が重要な課題として浮上しており、HITL はそれに対する理論的かつ実践的な対応として位置づけられている。なかでも、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。AI は、訓練データのバイアスやモデル設計上の制約に起因するバイアスを内包しやすく、過去の経済行動や制度に内在する構造的なバイアスを再生産する傾向がある。他方、人間の意思決定もまた、行動経済学が明らかにしてきたとおり、ヒューリスティクスや感情に基づく認知バイアスの影響を受ける。これら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、合理的かつ望ましい意思決定を実現するために不可欠である。したがって、HITL に基づく意思決定システムの設計においては、アルゴリズム・バイアスと認知バイアスの共通点および相違点を、理論的かつ実証的に検討する必要がある。こうした比較分析は、人間の判断力と AI の推論能力を相補的に活用し、信頼性の高い意思決定支援システムの構築に資する理論的基盤を提供する。

このような分析を進めるにあたっては、まず人間の意思決定においてどのようなバイアスが体系的に生じるのかを理解することが重要である。人間はしばしば、合理的な効用最大化モデルから逸脱する行動をとる。特に、危険や不確実性のある状況下においては、確実性効果や損失回避性といったバイアスが一貫して観察されてきた。これらの傾向は、Kahneman and Tversky (1979) のプロスペクト理論によって理論的に整理されており、その後、多くの実証研究が人間の非合理的な選好や判断のパターンを明らかにしてきた。たとえば、Falk et al. (2018) は 76 カ国・約 8 万人を対象とする国際調査を通じて、認知バイアスが国や文化だけでなく、年齢・性別・教育水準といった個人属性によっても異なることを示している。これらの知見は、HITL システムにおける人間の役割を設計するうえで、基礎的な理解を提供するものである。

近年、ChatGPT をはじめとする生成 AI (Generative AI) の発展に伴い、人間に見られる認知バイアスが AI にも現れるのかという問いが注目を集めている。生成 AI は、大規模な人間由来の言語データをもとに訓練されており、その過程で人間と類似した傾向やバイアスを内在的に獲得する可能性が指摘されている。仮に生成 AI が人間と同様のバイアスを再現しうるのであれば、社会調査、経済実験、政策シミュレーションといった分野において、有効な代替手段となる可能性を持つ。他方で、バイアスの増幅や出力の予測困難性といった新たなリスクも伴うことから、生成 AI の判断傾向の特性と

限界を明らかにすることは、現在の重要な研究課題である。こうした背景のもと、人間と生成 AI の判断傾向における類似点および相違点を明らかにしようとする比較研究が、近年急速に増加している。

中でも特に関心を集めているのが、生成 AI に特定のペルソナを与えることで、より人間らしい出力が得られるかという点である。ここでいうペルソナとは、AI エージェントに仮想的な個人属性(たとえば年齢、性別、教育水準、所得、文化的背景など)を付与することで、応答に一貫性や個性を持たせるための手法を指す。多くの先行研究では、このような属性情報をプロンプトに明示的に組み込むアプローチが採用されており、その結果として出力の現実性や一貫性が向上することが報告されている。たとえば、Park et al. (2024) は、米国人 1,052 名の詳細なインタビューデータをもとに、エージェントに年齢や政治的傾向といった属性を与えた際の行動が、実際の人間とどの程度一致するかを検証し、最大で 85% の一致率を示した。また、Jia et al. (2024) は、GPT モデルに人口統計情報を与えたうえで危険選好や損失回避傾向を測定し、性別や教育水準の違いが損失回避係数や確率重み付けに有意な影響を与えることを報告している。これらの知見は、生成 AI が人間に類似した行動パターンを再現しうる可能性を示しており、今後の実証的・応用的研究に対して重要な示唆を与えるものである。

損失および利得に関する傾向についても、多くの研究が生成 AI の特性を明らかにしている。共通して指摘されているのは、生成 AI は人間と同様に危険回避傾向を示す一方で、損失回避性については人間よりも弱い傾向が見られる点である。たとえば、Qiu et al. (2024) は、GPT-4 の損失回避係数が 1.09 であるのに対し、実際の消費者の係数は 2.56 と大きく、生成 AI が損失を相対的に過小評価する傾向にあることを報告している。また、Jia et al. (2024) も、生成 AI は全体として危険回避的であるものの、まれな事象に対してはその確率を過大に評価する傾向があると指摘している。モデル間の比較においては、Ross et al. (2024) が、GPT-4 は他のモデルと比べて最も安定しており、人間の選好構造に近い出力を示す一方で、Claude や Gemini では出力に大きなばらつきが見られることを示している。さらに、Chen et al. (2025) は、同一の選択課題に対してもプロンプトの文脈によって出力が大きく変化することを報告しており、生成 AI の選好構造には一貫性に乏しい側面があり、それがプロンプト設計に大きく依存している可能性を示唆している。

ファインチューニング (fine-tuning) 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を目的に応じて調整することが可能である。多くの研究においては、明示的に人間のデータを用いた再学習によるファインチューニングは行われていないが、一方で、HITL の一環として実施される人間のフィードバックに基づく強化学習 (Reinforcement Learning from Human Feedback: RLHF) が、モデルの出力傾向に大きな影響を及ぼす可能性が指摘されている。Santurkar et al. (2023) は、LLM の出力がリベラルかつ高学

歴・高所得層にバイアスを持つ傾向を示し、65 歳以上や低所得層、宗教的立場に基づく応答が十分に反映されていないことを報告している。また、RLHF のプロセスによって特定の価値観が強化され、バイアスが拡大する可能性があることも懸念されている。Hagendorff et al. (2023) も、RLHF がモデルの直感的な誤答を回避する能力や正確な回答を生成する能力を高める一方で、価値観の固定化を招くリスクを指摘している。他方、Park et al. (2024) は、実際の人間による回答データをプロンプトに反映させることで個別エージェントを構成する方法を採用しており、これは形式的なファインチューニングとは異なるものの、実質的には人間知識をモデルに注入するアプローチとして位置づけることができる。

このように、生成 AI は社会的属性や文脈設定を通じて人間に類似した行動傾向を示す一方で、その判断の忠実性や一貫性には依然として限界がある。また、RLHF によるフィードバック学習は出力の整合性や品質を高める可能性を有するものの、その実証的検証は依然として限定的である。したがって、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可能であるかを明らかにすることは、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、確実性効果および損失回避性に着目し、現実の人間と生成 AI (GPT-4) における意思決定バイアスの類似点および相違点を実証的に検討することを目的とする。具体的には、日本全国の成人 4,838 名を対象とした Web 調査を実施し、年齢・性別・世帯所得といった個人属性と回答傾向との関連性を分析する。さらに、得られた属性情報に基づいて仮想的なペルソナを構成し、異なる **temperature** 条件の下で GPT-4 に同一の設問を複数回提示する。そのうえで、生成 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における相違点および一致度を検証する。

本研究の中心的関心は、人間および生成 AI における危険選好と損失回避性の特徴を明らかにする点にあり、加えて、個人属性がこれらのバイアスに与える影響を比較検討することも重要な目的の一つである。これらの課題に対しては、人間と生成 AI の双方を対象とし、共通の枠組みに基づいた調査設計を通じて、バイアスの有無および傾向を測定・比較可能な手法を採用した。実証分析の結果、利得局面においては人間と生成 AI (GPT-4o) の選好傾向は概ね一致しており、いずれも危険回避的な行動を示すことが確認された。一方、損失局面では、生成 AI が人間よりも一貫して強い危険愛好性を示し、損失回避性に関して顕著な差が見られた。また、年齢や年収といった個人属性が意思決定に及ぼす影響については、生成 AI のほうがそれらの属性効果を人間よりも強く反映する傾向が認められた。さらに、**temperature** の設定によって出力の一貫性や属性効果の大きさが変化することも明らかとなった。これらの結果は、生成 AI が文脈や属性に基づくパターンを過剰に学習・反映する可能性を示唆しており、今後の応用においては、バイアスの補正や調整可能なチューニングの重要性が高まると

考えられる。

本論文の構成は以下のとおりである。第 2 章では、調査設計およびデータ収集の方法について述べる。第 3 章では、バイアスの推定手法を提示した上で、収集データに基づき人間および生成 AI の行動傾向を比較分析する。第 4 章では、推定結果の解釈と理論的含意について考察し、第 5 章では本研究の総括および今後の課題を論じる。

2. 調査設計

本研究は、人間および生成 AI における危険選好と損失回避性の特徴を明らかにすることを目的とする。あわせて、各個人属性がこれらのバイアスに与える影響について比較検討を行う。そのためには、人間と生成 AI の双方に対して、バイアスの有無および傾向を測定可能な調査が必要である。以上を踏まえ、本節では、両者に対するバイアス調査の設計と分析手法について説明する。

2.1. Web 調査の設計

我々は 2024 年 12 月、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を対象に Web 調査を実施した。調査にあたっては、日本全国を 9 地域に区分し、各地域で性別および年代が均等になるように回答者を割り当てた。調査実施時には、「消費者の属性と嗜好の関係性を把握することを目的とした学術的調査」である旨を明示し、対象者には回答への報酬が支払われることを事前に通知した。

その結果、5,040 名から有効回答が得られたが、不完全な回答 202 名分を除外し、最終的に 4,838 名のデータを分析対象とした。調査項目には、回答者の性別、年齢、世帯収入、学歴、職業、居住地といった個人属性に加え、認知バイアスや心理特性に関する複数の設問を含めた。後者には、Big Five 性格特性、時間選好、最後通牒ゲーム、トロッコ問題などに関する項目が含まれる。本論文では、個人属性のうち性別・年齢・世帯年収を分析に用いる。また、心理特性に関しては、プロスペクト理論に基づく設問のみに着目して検討を行う。表 1 には、回答者の性別・年齢・世帯年収に関する記述統計を示している⁴。

<表 1: 記述統計>

⁴ 年収は「300 万円未満」「300～500 万円未満」「500～700 万円未満」「700～1000 万円未満」「1000～1500 万円未満」「1500 万円以上」の 6 区分で質問し、それぞれ階級値をとって「300 万円」「400 万円」「600 万円」「850 万円」「1250 万円」「1500 万円」に変換した。

2.2. プロスペクト理論に関する質問

ここでは、危険非回避性および損失回避性を測定するための質問票について説明する。Tversky and Kahneman(1988)は、スタンフォード大学およびブリティッシュコロンビア大学の学生を対象に、プロスペクト理論の中核をなす危険選好および損失回避性を検証する目的で、利得局面および損失局面それぞれに関する以下の選択課題を提示した。

PROBLEM 1. (n=126) Assume yourself richer by \$300 than you are today. You are offered a choice between

- A. A sure gain of \$100, or
- B. A 50% chance to gain \$200, and A 50% chance to lose \$0.

PROBLEM 2. (n=128) Assume yourself richer by \$500 than you are today. You are offered a choice between

- A. A sure loss of \$100, or
- B. A 50% chance to lose \$200, and A 50% chance to lose \$0.

この研究では、利得局面と損失局面において、期待値が等しい2つの選択肢(AとB)を提示し、被験者の意思決定傾向を観察している。利得局面に対応する PROBLEM1 では、72%の被験者が選択肢 A(確実な利得)を選択し、28%が選択肢 B(高額利得の確率選択)を選んだ。この結果は、多くの被験者がより大きな利得の可能性よりも確実性を優先する傾向を持つことを示しており、利得状況における危険回避的な行動傾向がうかがえる。一方、損失局面に対応する PROBLEM2 では、36%の被験者が選択肢 A(確実な損失)を、64%が選択肢 B(損失回避の可能性を含む確率選択)を選択した。この結果は、多くの被験者が損失を完全に回避できる可能性に期待し、損失の拡大というリスクを受け入れたことを意味しており、損失状況における危険愛好的傾向を示している。

さらに注目すべきは、現在の富の状態を基準とした場合、PROBLEM1 および PROBLEM2 における選択肢 A は金額・確率の構造が同一であるにもかかわらず、選択傾向が明確に異なっている点である。このような選好の非対称性は、利得から得られる満足よりも損失を回避することによる満足の方が大きいという心理的傾向、すなわち損失回避性の存在を示唆している⁵。この現象は、従来の期待効用理論が前提とす

⁵ 同様の傾向は、Kahneman and Tversky (1979) においても報告されている。

る合理的意思決定とは異なる行動パターンを浮き彫りにしており、人間の判断が文脈や枠組みに大きく依存することを示している。

本調査では、Tversky and Kahneman (1988) の設問形式に依拠したうえで、以下のような設問を提示した。

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

なお、調査対象者が日本国内在住者であることから、設問は日本語で行い、金額を 1 ドル = 100 円で換算した円表示で提示した。

2.3. 生成 AI への調査とバイアスの推定

本研究の中核的な要素は、Web 調査で得られた実際の人間の属性データをもとに「ペルソナ」を構成し、それを生成 AI に実装したうえで、Web 調査と全く同一の質問を提示し、応答を取得する点にある。使用した生成 AI は、OpenAI 社の GPT-4o であり、一般的な ChatGPT サービスとは異なり、OpenAI API を通じてアクセスされるモデルである。この API を活用することで、Python コードを用いた大量かつ自動的な応答取得

同論文における設問の構成および結果は以下の通りである。PROBLEM1 では、選択肢 A および B を選択した被験者の割合はそれぞれ 16%、84%であった。一方、PROBLEM2 では、選択肢 C および D を選択した被験者の割合はそれぞれ 69%、31%であった。

PROBLEM 1. In addition to whatever you own, you have been given 1,000. You are now asked to choose between A: (\$1,000,.50), and B: (\$500).

PROBLEM 2. In addition to whatever you own, you have been given 2,000. You are now asked to choose between C: (\$-1,000,.50), and D: (\$-500).

が可能となった⁶。

生成 AI に対する質問提示から応答取得までのプロセスは、以下の手順で構成される。まず、Web 調査で得られた 4,838 名分のデータから、性別・年齢・世帯年収の 3 属性を抽出し、csv 形式で保存した。次に、この csv ファイルをもとにした Python スクリプトを生成 AI に読み込ませた(図 1)。1~5 行目の "import" によるコードは、csv の入出力や OpenAI API との接続に必要なモジュールを読み込むものである。7 行目の "openai.api_key" は、API の利用に必要な認証情報を設定する。9~10 行目の "file_path" は、読み込む対象ファイルのパスを指定する(ファイル名は伏せ字とした)。12~24 行目の "def extract_reason_and_answer" は、応答に含まれる「理由」および「回答」を抽出する関数であり、明確な理由が生成されない場合には「不明」と出力する仕様となっている。26~53 行目の "def generate_reason_and_answer" は、各ペルソナの属性をもとに生成 AI に回答を生成させる処理である。55 行目の "df[["理由", "回答"]]" により、これらの処理を 4,838 件すべてに対して実行し、それ以降のコードでは生成された応答結果を保存する。

<図 1: Python スクリプト>

さらに、生成される応答のばらつきを制御するために、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⁷。以上の手順により、Web 調査と同一の属性(性別・年齢・世帯年収)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(図 2)。得られた応答に対しては、前節で用いたロジット・モデルと同様の推定手法を適用し、生成 AI における危険選好および損失回避性の傾向を分析した。

<図 2: OpenAI API による応答>

3. ロジット・モデルの推定結果

本節では、前節までに示した手法に基づき、個人属性にもとづく「ペルソナ」ごとのロ

⁶ 我々は GPT-4o 以外にも、OpenAI 社の GPT-3.5、Google 社の Gemini 2.0 Flash、DeepSeek 社の DeepSeek-R1 のチャットサービスにも同様の質問を行い、応答を得た。結果については 4 節に記述している。

⁷ temperature は最小値が 0、最大値が 2、デフォルト値が 1 であり、値が大きくなるほど、生成テキストのランダム性が増す。

ジット・モデルによる推定結果について説明する。あわせて、第 1 項では推定手法の概要を述べ、第 2 項では推定に先立って得られた Web 調査および生成 AI の応答に基づく選択傾向の全体像を概観する。

3.1. バイアスの推定手法

本項では、調査対象者の属性が危険選好および損失回避性に与える影響を明らかにするため、推定に用いる手法について説明する。設問では、回答者はいずれか一方を選択する形式となっており、「選択肢 1=100%の確率で 1 万円」「選択肢 2=50%の確率で 2 万円」のいずれかを選ぶ。したがって、回答は二値変数として扱うことが可能である。

本分析では、この二値の回答を被説明変数とし、性別・年齢・世帯年収の 3 つの属性を説明変数として、最尤法による回帰分析を行う。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・モデルによって推定を行うことが適切である。

$$\log \left(\frac{\Pr(Y_i = 1)}{\Pr(Y_i = 0)} \right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

ここで、 i は各回答者を表している。 Y_i は回答者 i の選択を表す二値変数であり、「選択肢 1=100%の確率で 1 万円」を選択した場合は 1 を、「選択肢 2=50%の確率で 2 万円」を選択した場合は 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。

3.2. 選択傾向の概要

表 2 は、Web 調査の回答者(すなわち人間)および生成 AI による各質問への選択結果を示している。各質問に対して、「選択肢 1(100%の確率で 1 万円を得る、または失う)」を選択した場合を 1、それ以外の選択肢(選択肢 2)を選択した場合を 0 とするダミー変数を定義し、その平均値を求めることで、選択肢 1 の平均選択率を算出している。

まず、人間を対象とした Web 調査の結果を確認する。質問 1(利得局面)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、選択肢 2(50%の確率で 2 万円を得るが、50%の確率で何も得られない)を選択した割合(11.6%)を大きく上回った。この結果は、利得局面における危険回避的傾向を示唆している。一方、質問 2

(損失局面)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%、選択肢2(50%の確率で2万円を失うが、50%の確率で何も失わない)を選択した割合が42.3%であり、危険愛好的な選択も一定程度見られたが、明確な傾向とは言いがたい。

次に、 $temperature=1.0$ (デフォルト値)における生成AIの応答結果を示す。質問1では、選択肢1(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。一方、質問2では、選択肢1(確実な損失)を選んだ割合は11.0%にとどまり、多くの応答が選択肢2(確率的な損失)を選択しており、損失局面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における選択肢1の選択率はそれぞれ88.4%および91.1%であり、統計的に有意な差が認められた(p 値 <0.01)。ただし、両者ともに約9割の高い選択率を示しており、利得局面における危険回避性が共通して確認される。一方、質問2では、生成AIにおいて危険愛好性が明確に観察されるのに対し、Web調査の人間の回答ではその傾向は明確でなかった。この違いは、損失局面における選好構造が人間とAIの間で異なる可能性を示唆している。

また、質問1と質問2における選択肢1(確実な選択)の選択率を比較すると、生成AI($temperature=1.0$)では91.1%および11.0%であり、損失局面においてより強い損失回避性がうかがえる。一方、Web調査では88.4%および57.7%であり、生成AIと比較して損失回避性は相対的に弱い、その存在自体は確認される。

最後に、 $temperature$ の設定による生成AIの応答傾向の変化について述べる。質問1では、選択肢1の選択率が $temperature$ の設定値に応じて91.1%(1.0)、88.9%(0.5)、89.8%(0.0)といずれも高水準で推移しており、利得局面での危険回避性が一貫して観察された。一方、質問2における選択肢1の選択率は、 $temperature$ が高い順に11.0%(1.0)、3.0%(0.5)、0.0%(0.0)であり、 $temperature$ の低下にともなって危険愛好性が強まる傾向が確認された。特に $temperature=0.0$ の条件では、選択肢1を選択したのは「女性・53歳・年収300万円」の1件のみであり、この応答は外れ値とみなされる可能性が高い。したがって、以降の分析では当該データを除外する。

<表2: 選択結果>

3.3. ロジット・モデル限界効果

表3および表4は、各 $temperature$ の条件下における生成AIおよびWeb調査(人間回答)のデータを用いて、ロジット・モデルにより推定した限界効果および予測確率の平均値を示している。限界効果とは、ある説明変数が1単位変化した際に、 $Y_i=1$ となる確率がどの程度変化するかを示す指標であり、ロジット・モデルにおける係数の実

質的な解釈に用いられる。一方、予測確率の平均値とは、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 $Pr(Y_i = 1)$ を算出し、その平均値を求めたものである。この値は表 2 の選択率と一致しており、ロジット・モデルが実際の選択傾向を適切に予測できていることが確認される。

まず、人間を対象とした Web 調査の推定結果を概観する。質問 1 (利得局面) では、性別・年齢・年収のいずれの属性においても、有意水準 1% で統計的に有意な限界効果が確認された。具体的には、女性である場合に選択肢 1 を選ぶ確率が 4.1% 高くなり、年齢が 10 歳高い場合には 1.5% 高く、年収が 100 万円高い場合には 0.5% 低くなる傾向が見られた。一方、質問 2 (損失局面) では、年齢および年収においてのみ統計的に有意な限界効果が観察され、性別による影響は確認されなかった。具体的には、年齢が 10 歳高い場合に選択肢 1 を選ぶ確率が 2.1% 高くなり、年収が 100 万円高い場合には 0.8% 低くなるという結果が得られた⁸。

次に、`temperature=1.0` (デフォルト値) に設定した生成 AI の応答結果について検討する。この条件下では、質問 1 および質問 2 のいずれにおいても、すべての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。質問 1 では、女性である場合に選択肢 1 を選ぶ確率が 3.3% 高く、年齢が 10 歳高い場合には 5.7% 高く、年収が 100 万円高い場合には 2.7% 低くなる効果が示された。質問 2 では、女性である場合に選択確率が 2.8% 高く、年齢が 10 歳高い場合には 2.3% 高く、年収が 100 万円高い場合には 0.6% 低くなる傾向が確認された。

続いて、人間の Web 調査の結果と生成 AI (`temperature=1.0`) の応答結果を比較する。比較にあたっては、デルタ法により限界効果の標準誤差を算出し、両者の限界効果の差に関する検定を実施した。質問 1 における性別・年齢・年収の z 値は、それぞれ 0.70、-9.90、15.56 であり、性別については 5% 水準で有意な差は認められなかった。一方で、年齢および年収については、統計的に有意な差が確認された。質問 2 では、性別・年齢・年収の z 値がそれぞれ -2.04、-0.31、-0.71 となり、性別についてのみ有意な差が認められたが、年齢および年収については有意な差は見られなかった。以上の結果から、人間と AI の限界効果には部分的に差異が存在するものの、全体として統計的に有意な差があるとは言い難いことが示された。

最後に、`temperature` の違いによって生成 AI の応答結果がどのように変化するかを検討する。いずれの `temperature` 設定においても、質問 1 および質問 2 において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。質問 1 の年齢に関する限界効果は、人間の Web 調査の 0.015 と比べて生成 AI ではより大きく、`temperature` が低くなるにつれて限界効果が増大する傾向が見られた (`temperature=`

⁸ 推定に用いたロジット・モデルは、性別、年齢、年収の 3 属性に加えて、学歴 (大卒ダミー)、職業 (有職ダミー)、心理特性 (Big5) を説明変数とする 6 属性モデルへと拡張可能であり、その場合でも限界効果の傾向に大きな変化は見られない。

1.0、0.5、0.0 でそれぞれ 0.057、0.073、0.120)。一方、年収に関しては、Web 調査の限界効果(-0.005)と比較して生成 AI の値は小さくなる傾向があるものの、temperature との間に明確な関係性は見られなかった(-0.027、-0.033、-0.031)。性別についても、temperature の変化に伴う一貫した傾向は確認されなかった(Web 調査では 0.041、生成 AI では temperature=1.0、0.5、0.0 でそれぞれ 0.033、0.026、0.086)。

質問 2 に関しては、年齢についても一貫した temperature 依存の傾向は確認されず、人間の Web 調査の限界効果(0.021)に対し、生成 AI では temperature=1.0、0.5 でそれぞれ 0.023、0.008 となった。年収については、Web 調査の限界効果(-0.008)と比較して生成 AI はやや大きい傾向を示したが、temperature の変化による明確な影響は認められなかった(いずれも-0.006)。性別についても、temperature による限界効果の変化は確認されなかった(いずれも 0.028)。

<表 3:ロジット・モデル推定結果(利得:3 属性)>

<表 4:ロジット・モデル推定結果(損失:3 属性)>

3.4. 選択確率の図解

図 3 および図 4 は、それぞれ質問 1 および質問 2 における各属性の選択確率を示している。たとえば、図 3 の「利得 3 属性 性別」では、性別以外の 2 属性を平均値(表 1 に記載のとおり、年齢=44.9 歳、年収=593 万円)に固定し、性別のダミー変数が 0 の場合と 1 の場合における選択確率が図示されている。図 3 から読み取れる傾向として、まず「利得 3 属性 性別」では、 $x=0,1$ のいずれの点においても 4 本の折れ線の 95%信頼区間が重なっており、Web 調査結果と生成 AI の応答結果、また temperature の違いによる選択確率の差は確認されない。

一方、「利得 3 属性 年齢」および「利得 3 属性 年収」では、Web 調査結果を示す黒線の傾きが比較的緩やかであるのに対し、生成 AI の線はより急峻となっており、年齢や年収に応じた選択確率の変化を生成 AI がより強調していることが読み取れる。ただし、いずれのグラフにおいても平均値付近(年齢=4.49、年収=5.93)では、4 本の折れ線の 95%信頼区間が重なっており、選択確率の差は小さいことが確認できる。図 4 では、性別・年齢・年収のいずれの属性においても、y 切片に明確な差が認められ、Web 調査結果が 0.5~0.6 前後、生成 AI の結果が 0~0.1 前後に位置している。一方、各線の傾きについては、両者間で大きな違いは見られない。

さらに、両図の y 座標の水準を比較することで、損失回避性の傾向を視覚的に把握することができる。たとえば、図 3「利得 3 属性 選択確率 年齢」では、生成 AI (temperature=1.0)における 20 歳時点での選択確率がおよそ 0.7 であるのに対し、図

4「損失 3 属性 選択確率 年齢」では、同条件下での選択確率は約 0.05 にとどまっている。この大きな差は、第 2 節第 2 項で述べたとおり、生成 AI における強い損失回避性を反映していると考えられる。この図から読み取れる重要な点は、人間のデータでは損失回避性が相対的に弱く表れる一方で、生成 AI においてはその傾向が顕著に現れているという点である。

<図 3: 選択確率(利得:3 属性)>

<図 4: 選択確率(損失:3 属性)>

3.5. ロジット・モデルにおける的中率

表 5 は、各モデルによって予測された選択確率と実際の選択値との一致度(的中率)を示している。ここで「予測値」とは、予測確率 $\Pr(Y_i = 1)$ が 0.5 以上であれば 1、0.5 未満であれば 2 と分類した二値変数である。この分類に基づき、実際の回答との一致度を算出することで、モデルの適合度を評価している。人間の Web 調査結果における的中率は、質問 1 において 88.40%と高い値を示した一方、質問 2 では 57.59%にとどまった。これに対して、生成 AI の応答に基づく的中率は 88.98%から 96.96%の範囲にあり、全体として Web 調査よりも高い適合度を示している。また、Web 調査における質問 1 および生成 AI による質問 2 の結果では、モデルの予測が一方の選択肢に偏る傾向も確認された。

<表 5: ロジット・モデルの的中率>

4. 考察

本研究により、OpenAI 社が提供する GPT-4o が危険選好および損失回避性をどのように学習しているのか、その一端を明らかにすることができた。質問 1 においては、人間の Web 調査および生成 AI の応答結果のいずれにおいても、選択肢 1 の選択率は 9 割前後であり、利得局面における危険回避的傾向を概ね忠実に再現しているといえる。属性ごとの限界効果を比較すると、性別については人間と生成 AI の間に大きな差は見られなかったが、年齢および年収に関しては、いずれも正の限界効果が観察され、生成 AI の係数の方が大きくなる傾向が確認された。これは、生成 AI が属性による影響を人間よりも強調する傾向を有していることを示唆している。とりわけ、temperature を 0.0 に設定した場合には、他の設定条件と比べて係数の絶対値が大き

くなる傾向が顕著であり、出力のランダム性が抑制されることによって、実際には選択にばらつきが見られる人間よりも、より明確な選好パターンが反映された結果と解釈できる。

一方、損失局面においては、生成 AI は人間よりも顕著な危険回避性を示した。質問 2 に関する人間の Web 調査では、選択肢 1 の選択率は 57.7% であり、Tversky and Kahneman (1988) の報告する 36% と比較しても、一定の危険回避的傾向が確認される。これに対し、生成 AI の選択率は、temperature = 1.0 および 0.5 の場合でそれぞれ 11.0%、3.0% にとどまり、より強い危険回避傾向を示した。さらに、temperature = 0.0 では、1 つのサンプルを除き、すべてのサンプルが選択肢 2 のみを選択する結果となり、危険を極端に忌避する傾向が観察された。

損失回避性についても、同様の傾向が確認された。質問 1 と質問 2 における選択肢 1 の選択率を比較すると、Tversky and Kahneman (1988) では 72% 対 36%、人間の Web 調査では 88% 対 58%、生成 AI ではおおむね 90% 対 10% となっており、生成 AI の方が人間よりも損失をより強く忌避する傾向を有していることが示されている。さらに、表 5 に示された的中率の比較においても、Web 調査では予測値が 1 であったサンプルが 98.04% だったのに対し、生成 AI ではすべてのサンプルが予測値 2 となっており、危険選好および損失回避に関する傾向に明確な差が存在することがうかがえる。

ただし、属性ごとの限界効果に着目すれば、生成 AI はむしろ属性間の違いを適切に表現しているとも評価できる。さらに、第 2 節第 3 項で述べたとおり、GPT-4o 以外の LLM——すなわち OpenAI 社の GPT-3.5、Google 社の Gemini 2.0 Flash、DeepSeek 社の DeepSeek-R1——についても、同様の質問を提示したところ、各モデル間で明確な違いが見られた。GPT-3.5 には危険選好に関する理解の不十分さが認められ、Gemini 2.0 Flash は設問の意図を正確に把握していない応答が目立った。一方、DeepSeek-R1 は、高い精度で危険選好や損失回避性に関する研究結果を反映し、GPT-4o に類似した応答を示した。以上の結果から、GPT シリーズにおいては 3.5 から 4o へのバージョンアップによって学習内容が深化していること、Gemini における当該特性の学習は現時点では限定的であること、そして DeepSeek は GPT-4o と同等の水準で当該特性を学習している可能性があることが示唆される。

5. 結論

本研究は、危険選好および損失回避性に関する人間と生成 AI (GPT-4o) の意思決定傾向を比較し、特に属性情報に基づくペルソナ設定が AI の応答に与える影響を検討した。その結果、利得局面においては人間と生成 AI のあいだに大きな差異は認められず、いずれも危険回避的な傾向を示した。一方、損失局面では、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、年齢や所得といった個人属性の影響については、生成 AI の方が人間よりも効果を強調する傾向が観察され、これは選好のバイアスを増幅する方向に作用していると解釈される。さらに、*temperature* の設定が出力傾向に与える影響も顕著であり、ランダム性の低い条件下では、より一貫性のある応答が得られる一方で、現実の人間の選好とは乖離した出力が観察された。他モデルとの比較においては、GPT-4o は特に安定的であり、人間の選好傾向に最も近い出力を示したが、損失回避性に関しては依然として明確な差異が残された。以上の結果を踏まえ、今後の課題としては、人間の選好をより忠実に再現するためのファインチューニング手法の確立と、属性情報に基づくバイアスの過剰な強調を抑制する仕組みの導入が挙げられる。また、本研究では認知バイアスの一例として危険回避性に着目したが、今後は時間選好や現在性など、他のバイアスについても同様の枠組みによる分析を進めていく予定である。

参考文献

1. Chen, Y., Kirshner, S. N., Ovchinnikov, A., Andiappan, M., & Jenkin, T. (2025). A manager and an AI walk into a bar: does ChatGPT make biased decisions like we do?. *Manufacturing & Service Operations Management*.
2. Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *The quarterly journal of economics*, 133(4), 1645-1692.
3. Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science*, 3(10), 833-838.
4. Jia, J., Yuan, Z., Pan, J., McNamara, P., & Chen, D. (2024). Decision-making behavior evaluation framework for llms under uncertain context. *Advances in Neural Information Processing Systems*, 37, 113360-113382.
5. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 363-391.
6. Park, J. S., Zou, C. Q., Shaw, A., Hill, B. M., Cai, C., Morris, M. R., ... & Bernstein, M. S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint*, arXiv:2411.10109.
7. Qiu, L., Singh, P. V., & Srinivasan, K. (2023). Consumer Risk Preferences Elicitation From Large Language Models. Available at SSRN 4526072.
8. Ross, J., Kim, Y., & Lo, A. W. (2024). Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint*, arXiv:2408.02784.
9. Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? *In International Conference on Machine Learning*, pp. 29971-30004.
10. Tversky, A., & Kahneman, D. (1988). Rational choice and the framing of decisions. In D. E. Bell, H. Raiffa, & A. Tversky (Eds.). *In Decision making: Descriptive, normative, and prescriptive interactions*, pp. 167-192.

図表

表 1: 記述統計

属性	平均
女性ダミー	0.507 [0.500]
年齢(10歳)	4.49 [1.33]
年収(100万円)	5.93 [3.10]
Obs.	4,838

Notes: 表中の括弧内の数値は標準偏差を示している。

図 1: Python スクリプト

```
1 import time
2 import pandas as pd
3 import openai
4 import os
5 import re
6
7 openai.api_key = os.getenv("OPENAI_API_KEY")
8
9 file_path = r"C:\Users\***\***.csv"
10 df = pd.read_csv(file_path)
11
12 def extract_reason_and_answer(response_text):
13     match_reason = re.search(r"理由[:]\s*(.*)", response_text)
14     match_answer = re.search(r"回答[:]\s*[12])", response_text) # 「回答:1」をそのまま取得
15
16     if not match_reason:
17         lines = response_text.split("\n")
18         reason = lines[0].strip() if lines else "不明"
19     else:
20         reason = match_reason.group(1).strip()
21
22     answer = match_answer.group(1).strip()
23
24     return reason, answer
25
26 def generate_reason_and_answer(row):
27     question = f"""性別: {row['性別']}, 年齢: {row['年齢']}歳, 年収: {row['年収']}万円の人に以下の質問をします。
28     あなたは現在の富に上乗せて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？
29     1. 確実に10,000円もらえる
30     2. 50%の確率で20,000円もらえて、50%の確率で何ももらわない
31     回答は以下のフォーマットで厳密に記述してください：
32     - 理由: [あなたの選択の理由]
33     - 回答: [1 または 2 のみを出力]
34     """
35
36     for attempt in range(3):
37
38         time.sleep(1)
39         response = openai.chat.completions.create(
40             model="gpt-4o",
41             messages=[
42                 {"role": "system", "content": f"あなたは性別: {row['性別']}, 年齢: {row['年齢']}歳, 年収: {row['年収']}万円の人です。"},
43                 {"role": "user", "content": question}
44             ],
45             max_tokens=180,
46             temperature=0.0
47         )
48
49         full_response = response.choices[0].message.content.strip()
50         print(f"✅ API Response for row {row.name}: {full_response}")
51
52         reason, answer = extract_reason_and_answer(full_response)
53         return pd.Series([reason, answer])
54
55 df[["理由", "回答"]] = df.apply(generate_reason_and_answer, axis=1)
56
57 output_file_path = r"C:\Users\***\***.csv"
58 df.to_csv(output_file_path, index=False)
59
60 print(f"✅ 理由と回答を生成し、新しいファイルに保存しました: {output_file_path}")
```

Notes: 「あなたは性別: {row['性別']}, 年齢: {row['年齢']}歳, 年収: {row['年収']}万円の人です。」と生成 AI を役割付け、質問を行った。なお、本図は実際に使用したスクリプトより、生成 AI による応答を出力させるうえで重要な部分を抜粋したものである。

図 2: OpenAI API による応答

	年齢	年収	理由	回答
男性	26	400	現在の収入や年齢を考え、危険を冒すことよりも確実に利益を得たいという考えから、10,000 円を確実にもらえる方が安心できるため。	回答: 1
女性	39	850	確実に 10,000 円をもらうことで危険がなく、手元の資金を安定して増やすことができるため。	回答: 1
男性	38	400	確実に 10,000 円をもらう方が危険が少なく、収入が限られている中で安定して利益を得ることができるため。	回答: 1
男性	22	850	現在の収入や年齢を考えると、危険を取る余裕があり、期待値を重視して選びたいと思います。2 の選択肢では期待値が 10,000 円なので危険を取ります。	回答: 2
男性	61	400	確実に 10,000 円をもらうことで、危険を避けながら少額でも追加の収入を得ることができるため。	回答: 1
女性	41	850	私は安定性を重視するため、確実に手に入る価値がある選択肢を好みます。確実に 10,000 円もらえる方が危険がないため、選びます。	回答: 1
男性	47	300	現在の収入を考慮すると、お金を確実に増やす方が危険が少なく、家計にありがたいと感じます。安定した収入を優先したいので、確実に 10,000 円もらえる選択肢を選びます。	回答: 1
男性	38	850	現在の年収や経済状況から考慮すると、大きな危険を取らずに確実に収入を増やすことを選びたいと考えます。10,000 円が確実に手に入ることで、ますます生活の余裕を確保できるためです。	回答: 1
男性	55	300	収入が限られているため、確実に 10,000 円を得ることで、すぐに安心感を得たいと考えます。危険を避けて確実な選択をすることがより自分にとって良いと思うからです。	回答: 1

Notes: 2 節 3 項の作業を通じて得られた応答の抜粋(図 2 は質問 1 に対する応答、設定は temperature=1.0)は、csv 形式で出力される。

表 2: 選択結果

質問 1	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
Mean	0.898	0.889	0.911	0.884
Std. Dev.	0.302	0.314	0.285	0.320
Obs	4,838	4,838	4,838	4,838

質問 2	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
Mean	0.000	0.030	0.110	0.577
Std. Dev.	0.014	0.172	0.313	0.494
Obs	4,838	4,838	4,838	4,838

表 3:ロジット・モデル推定結果(利得:3 属性)

質問 1		AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
限界効果	女性ダミー	0.086*** (0.006)	0.026*** (0.007)	0.033*** (0.007)	0.041*** (0.009)
	年齢 (10 歳)	0.120*** (0.004)	0.073*** (0.003)	0.057*** (0.003)	0.015*** (0.003)
	所得 (100 万円)	-0.031*** (0.001)	-0.033*** (0.001)	-0.027*** (0.001)	-0.005*** (0.001)
obs.		4,838	4,838	4,838	4,838
R^2		0.6355	0.3932	0.3472	0.0140
予測確率		0.898 (0.235)	0.889 (0.191)	0.911 (0.158)	0.884 (0.032)

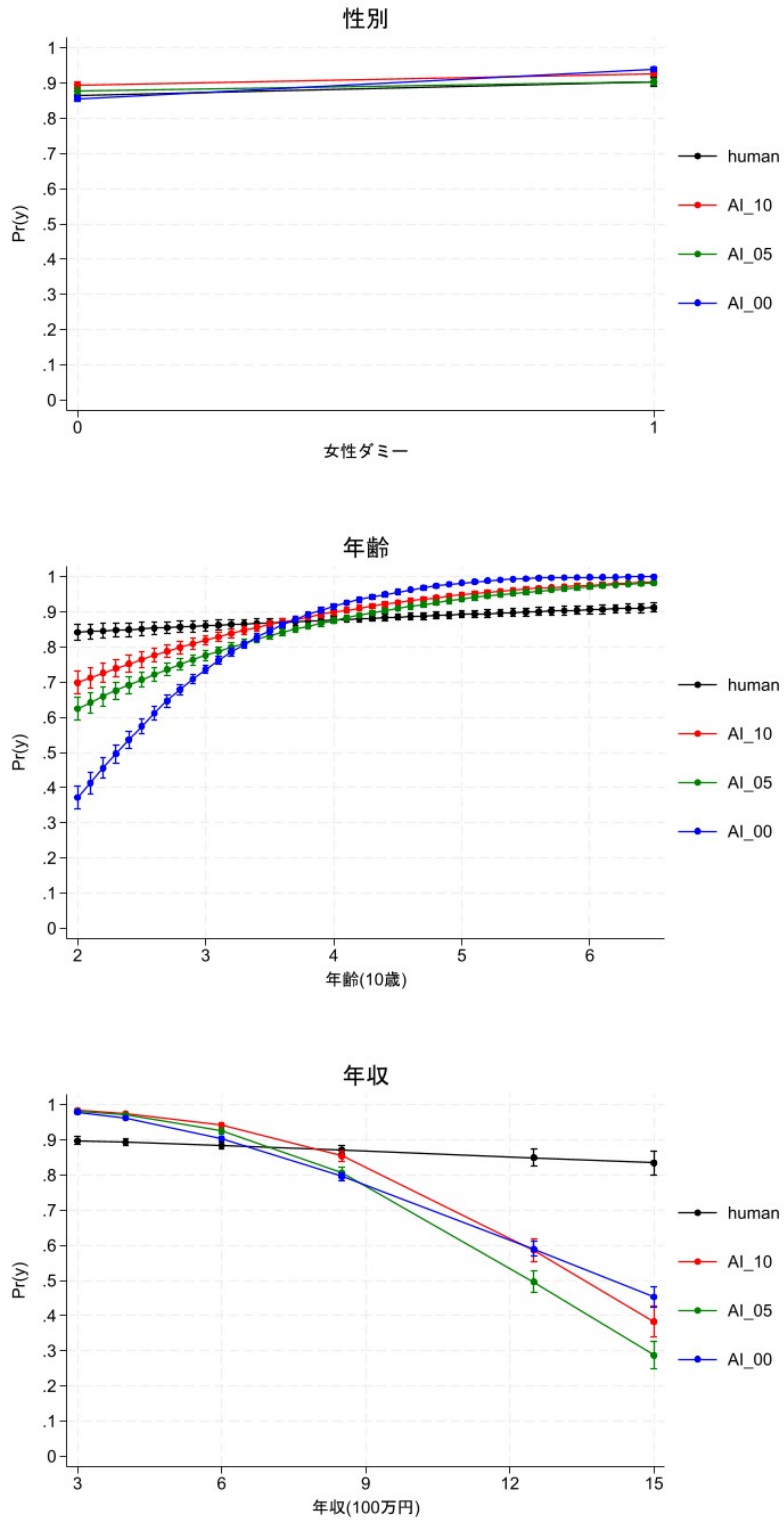
Notes: 表中の 2 行目から 4 行目の数字はロジット・モデルによる各属性の限界効果を示している。表中の括弧内の数字は標準誤差を表している。また、***は有意水準 1%で統計的に有意であることを示している。7 行目には、ロジット・モデルの推定結果に基づいた予測確率の平均値を示している。

表 4:ロジット・モデル推定結果(損失:3 属性)

質問 1		AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
限界効果	女性ダミー	-	0.028*** (0.006)	0.028*** (0.009)	-0.006 (0.014)
	年齢 (10 歳)	-	0.008*** (0.002)	0.023*** (0.004)	0.021*** (0.005)
	所得 (100 万円)	-	-0.006*** (0.001)	-0.006*** (0.002)	-0.008*** (0.002)
obs.		-	4,838	4,838	4,838
R^2		-	0.0644	0.0204	0.0038
予測確率		-	0.030 (0.024)	0.110 (0.038)	0.577 (0.035)

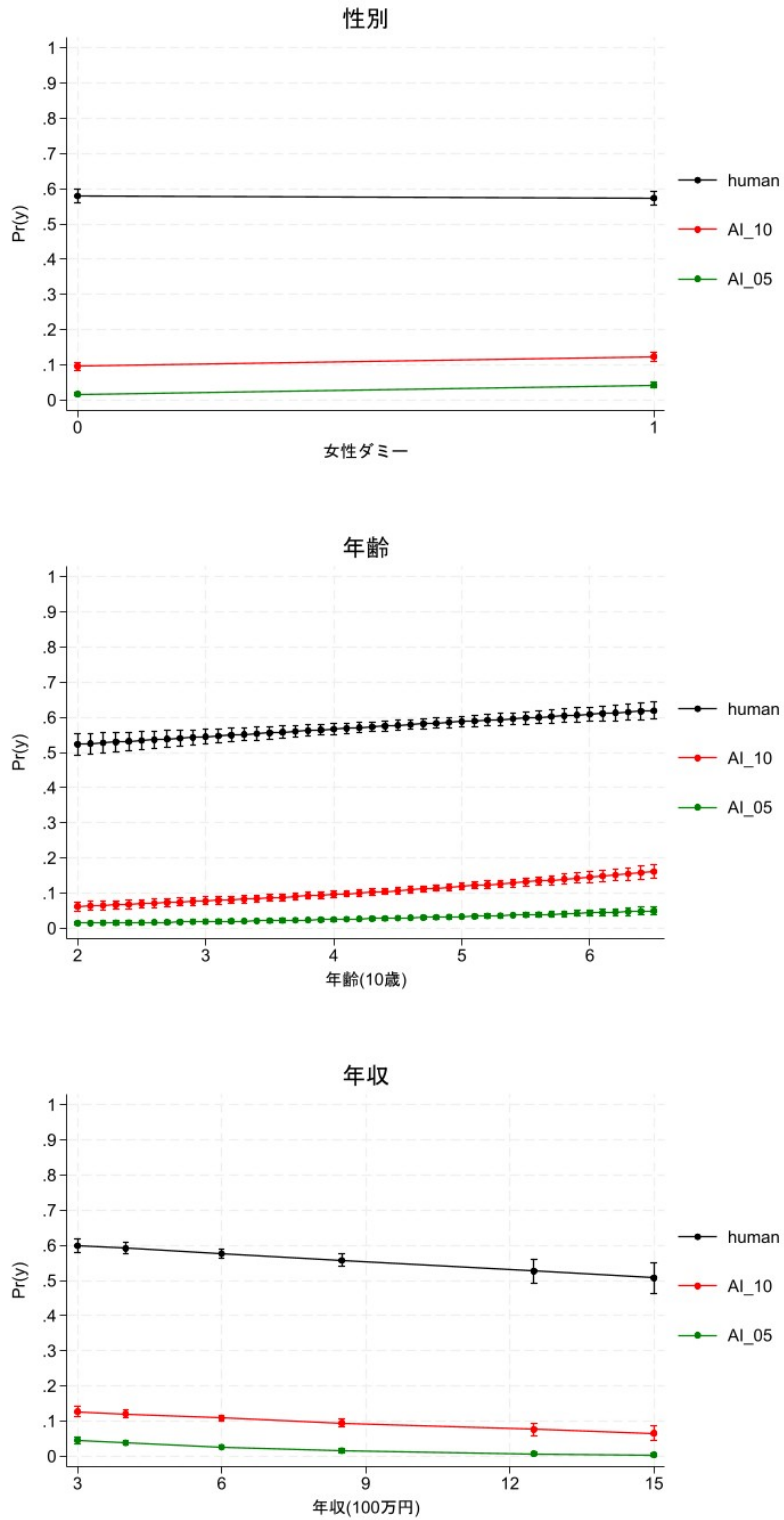
Notes: 表中の 2 行目から 4 行目の数字はロジット・モデルによる各属性の限界効果を示している。表中の括弧内の数字は標準誤差を表している。また、***は有意水準 1%で統計的に有意であることを示している。7 行目には、ロジット・モデルの推定結果に基づいた予測確率の平均値を示している。

図 3: 選択確率 (利得:3 属性)



Notes: AI_10 とは、AI•temperature 1.0 で尋ねた場合のことを指している。AI_05、AI_00 も同様。また、図中の各点から生えているひげは、95%信頼区間を表す。

図 4: 選択確率 (損失: 3 属性)



Notes: AI_10 とは、AI・temperature 1.0 で尋ねた場合のことを指している。AI_05、AI_00 も同様。また、図中の各点から生えているひげは、95%信頼区間を表す。

表 3:ロジット・モデルの的中率

質問 1	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
予測値,観測値				
(2, 2)	6.39% (n=309)	4.24% (n=205)	2.34% (n=113)	0% (n=0)
(2, 1)	1.53% (n=74)	2.00% (n=97)	1.34% (n=65)	0% (n=0)
(1, 2)	3.78% (n=183)	6.84% (n=331)	6.57% (n=318)	11.60% (n=561)
(1,1)	88.30% (n=4,272)	86.92% (n=4,205)	89.75% (n=4,342)	88.40% (n=4,277)
的中率	94.69%	91.15%	92.08%	88.40%

質問 2	AI temperature 0.0	AI temperature 0.5	AI temperature 1.0	人間
予測値,観測値				
(2, 2)	-	96.96% (n=4,691)	88.98% (n=4,305)	0.93% (n=45)
(2, 1)	-	3.04% (n=147)	11.02% (n=533)	1.03% (n=50)
(1, 2)	-	0% (n=0)	0% (n=0)	41.38% (n=2,002)
(1,1)	-	0% (n=0)	0% (n=0)	56.66% (n=2,741)
的中率	-	96.96%	88.98%	57.59%