



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

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 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

2026 年 3 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

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

2026年3月18日

本稿は2025年4月11日付の中間報告版にチューニング分析を加えた2025年6月25日版を
基にした最終報告の修正版である。

岩本涼太* 石原卓典† 依田高典‡

要旨

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・年収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を再訓練することにより、学習後のAIは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示された。

JEL 分類番号:D91, C91

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

* 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(Rahwan 2018, Rahwan et al. 2019)。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。この比較分析は、人間の判断力とAIの推論能力を相補的に活用することで、信頼性の高い意思決定支援のための理論的基盤を提供する⁴。

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、損失回避性(loss aversion)といったバイアスが傾向と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、初期の損失回避性を検証した研究としてはTversky and Kahneman(1992)がある。彼らの研究では損失回避度が $\lambda = 2.25$ 程度であると推定されており、人間は利得の2倍程度損失を感じ、その結果損失を回避すると言われている。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、損失回避性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。またBrown et al.(2024)では経済学や心理学などの分野における150本の論文から得られた推定値をもとに、損失回避性についてのメタ分析を実施しており、中央値で $\lambda = 1.69$ 程度の損失回避度が報告されている。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、

⁴ 本研究は、Human-in-the-Loop 機械学習のプロセスを、一気通貫した定量的メトリクスに基づき、行動経済学において実装した先進的事例として位置付けられる。本研究とHITLの関係については、Appendixを参照されたい。

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。このような背景のもと、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。

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(risk preferences)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。

利得および損失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 は人間と部分的に異なる危険選好を見せる点である。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、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、危険選好の感応度逡減(diminishing sensitivity)と損失回避性と関連が深い「参照点依存(reference dependence)」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設計し、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。

ここでは、参照点依存は、ゼロ円を参照点として、利得領域における危険回避的行動および損失領域における危険選好的行動として観察される⁵。本研究の実証分析によれば、利得領域においては、人間と GPT-4o の危険回避的傾向におおむね共通性が認められる。一方、損失領域においては、GPT-4o は人間と比べて有意に強い危険選好的傾向を示しており、これは参照点依存の強度または関数形の違いに起因するリスク選好の乖離を示唆している。さらに、意思決定における人口統計学的手がかりへの感応度に関して、GPT-4o は人間よりも高い反応性を示し、年齢や所得といった個人属性が AI の応答に与える影響は、人間の選択におけるそれよりも大きい傾向が認められた。

さらに、人間の参照点依存に関する Web 調査データを用いて AI をファインチューニングした結果、AI は選択傾向や属性ごとの限界効果において人間の行動様式に近づいたことが確認された。とりわけ損失においては、学習前に比して人間に類似した選好パターンを示すようになり、学習後の AI は人間と学習前 AI の中間的な行動特性を備えるに至った。ただし、正解率や分布の一致度の面では、一部に過学習や乖離も見られた。分布間の類似性を評価するためにワッサースタイン距離を用いた分析においても、学習後 AI は学習前 AI に比べて人間の選択分布に統計的に近接しており、ファインチューニングが行動分布の収束に寄与することが示された。

⁵ 本研究は利得領域と損失領域における価値関数の差異を直接推定するものではない。むしろ、損失回避性と感応度逡減の主要な行動的含意、すなわち両領域間における危険選好の参照点依存に焦点を当てており、これは人間および AI エージェントの双方において観察される現象である。

本論文の構成は以下のとおりである。第2章では、生成AIと人間の比較研究についての先行研究を紹介する。次に第3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4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と学習前AIの行動傾向を比較分析する。第5章では、人間の選択傾向をもとにファインチューニングを行い、学習後AIと人間の行動傾向を比較分析する。第6章では、人間と学習前AI、学習後AIのバイアス傾向を比較し、それらの関係性についてまとめる。第7章では、分析結果をもとに考察を行い、最後に第8章で本研究の総括および今後の課題を論じる。

2. 先行研究

本節では、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を概観する⁷。表1は、本節で取り上げる各研究について、以下の観点から整理したものである。(1)対象としているバイアスや実験の内容、(2)使用されている生成AIの種類・バージョン、(3)研究対象およびサンプルサイズ、(4)LLMによる人間行動の平均的な再現性、(5)社会経済的属性で条件つけた際の再現性、(6)ファインチューニングなど人間からのフィードバックの有無、および(7)人間とLLMの回答分布の違いを評価するために用いられたメトリクスである。

<表1:生成AIと人間の意思決定・バイアス比較を行っている先行研究>

ここでは、経済実験を通じて人間と大規模言語モデル(LLM)の意思決定を比較した研究を紹介する。まず、利他性に代表される社会的選好や戦略的相互作用を扱うゲーム理論実験に焦点を当てる。人間との直接的な比較は行っていないものの、LLM同士を経済ゲームに参加させて行動を評価した研究として、Akata et al.(2023)、Guo(2023)、Tsuchihashi(2023)、Lorè and Heydari(2024)が挙げられる。Akata et al.(2023)は、GPT-3、GPT-3.5、GPT-4を用いて、異なるペイオフ行列を持つ囚人のジレンマや男女の争いゲームにおける繰り返しゲームを実施し、協力率および選択行動を測定した。GPT-4は人間に類似した協力行動を示したが、暗黙の協調が求められる場面では協調に失敗する傾向が観察された。加えて、プロンプトに「協調的ペルソナ」を導入することで協調成

⁷ 近年では、大規模言語モデル(LLM)の研究応用に関する展望論文やレビュー論文も複数登場しており、それらはLLMの可能性と限界の両面を理論的に整理している。Demszky et al.(2023)は、心理学分野におけるLLMの活用法を体系的に整理し、とりわけ文化的バイアスや代表性の低いグループが欠如している場合に、AIの出力に偏りが生じるリスクについて言及している。Hagendorff(2023)は「機械心理学(Machine Psychology)」という概念を提唱し、LLMを心理学的実験の仮想的な参加者とみなす枠組みを提示するとともに、こうしたアプローチに内在する倫理的問題への注意を促している。また、Sarstedt et al.(2024)は、いわゆる「シリコンサンプル」と人間の回答を比較した研究のレビューを通じて、LLMが消費者行動やマーケティング調査への応用可能性を持つかを検討している。彼らは、ブランド評価やフレーミング効果といった特定の課題においては高い再現性が認められる一方で、選好形成や行動経済学的意思決定タスクの再現には限界があることを指摘している。これらのレビュー研究は、LLMを社会科学研究に導入する際の指針を与えるとともに、今後の活用領域の選定や倫理的ガバナンスにおいても重要な視座を提供している。

功率が向上することが示された。Guo(2023)は GPT-4 に「公平」「利己」といったペルソナを付与し、最後通牒ゲームおよび囚人のジレンマを複数回実施した。公平ペルソナでは、人間の平均に近い提案額と拒否閾値が示され、利己ペルソナでは低提案・高受容の選択傾向が観察された。Lorè and Heydari(2024)は、4種類の二人ゲームを5つの文脈条件で出題し、GPT-3.5、GPT-4、LLaMa-2の行動を比較した。GPT-4はゲーム構造に依存して合理的な戦略を選ぶ傾向があった一方、GPT-3.5およびLLaMa-2は文脈的な影響を受けやすいことが報告されている。Tsuchihashi(2023)はGPT-3.5に封印入札オークションへの参加を促し、LLMの入札行動を評価した。一次価格オークションでは人間と同様に過大入札の傾向がみられ、二次価格オークションでは理論に沿った真値入札が行われた。さらに、「学生ペルソナ」を与えることで、FPAにおいて理論的により妥当な入札を行い、SPAでは過小入札となる傾向が確認された。

人間の行動との比較に基づき、社会的選好に関する実験を行った研究には、Horton(2023)、Brookins and DeBacker(2024)、Mei et al.(2024)、Xie et al.(2024)がある。Horton(2023)は、LLMをホモ・シリカス(*Homo Silicus*)という仮想的な経済主体と見なし、GPT-3に対して価格の公正性評価、現状維持バイアス、最低賃金の影響など、標準的な行動経済実験課題を提示した。その結果、GPT-3は既存の人間実験と同様のバイアスを示すことが確認され、さらに利己的・公平・効率志向といったペルソナ設定によって意思決定の傾向が変化することが示された。Brookins and DeBacker(2024)はGPT-3.5に独裁者ゲームと囚人のジレンマを複数回プレイさせ、公平性および協調行動の傾向を分析した。その結果、AIは人間よりも高い利他性を示す傾向があり、特に効率性を重視する文脈においては人間と類似した行動がみられた。Mei et al.(2024)は、公開データに基づき、6種類のゲームにおける約9万件の人間による選択データをもとに、GPT-4およびGPT-3.5の出力との類似性を、チューリングテストの枠組みで評価した。GPT-4の出力は平均的には人間の回答と類似しており、高い再現性が報告された。他方で、AIの出力は中央値に集中する傾向があり、極端な選択を避ける傾向が確認されている。Xie et al.(2024)は、GPT-4oを含む5種類のモデルに対して独裁者ゲームや最後通牒ゲームなどを実施し、LLMと人間の回答分布の違いをWasserstein距離により評価した。全体として、LLMは人間よりも公平で協調的な選択を行う傾向があり、分布的にはGPT-4oが人間の回答に最も近いと報告されている。

損失回避や危険選好に関連する実験研究としては、Jia et al.(2024)、Qiu et al.(2024)、Ross et al.(2024)、Chen et al.(2025)、Macmillan-Scott and Musolesi(2024)などがある。Jia et al.(2024)は、損失回避と確率加重に関する選択課題をGPT-4やClaude、Geminiに提示し、回答傾向を評価した。すべてのモデルがリスク回避的な傾向を示し、特にGPT-4は人間と同様の損失回避行動を示した。社会経済属性によるプロンプト条件の違いにより、出力結果が大きく変化することも報告されている。Ross et al.(2024)は、GPT-4を含む12のLLMを対象に、不平等回避、損失回避、時間割引といったバイアスを包括的に分析した。LLMは、他者への罪悪感が強く嫉妬は弱い傾向を示し、利益については合理的に反応する一方、損失については非合理的な傾向が観察された。また、時間割引に関しては、人間よりも強い割引傾向が確認された。Chen et al.(2025)は、リスク回避を含む18種類の認知バイアスを検証し、GPT-3.5では利得に対してリスク愛好的、損失に対してリスク回避的な

傾向がみられたのに対し、GPT-4 ではリスク回避的であり、フレーミング効果の影響が相対的に小さいことが報告された。Macmillan-Scott and Musolesi(2024)は、条件付き確率の誤謬やギャンブラーの誤謬など 12 の認知バイアス課題を提示し、GPT-3.5 は人間に類似したバイアスを多く示したが、GPT-4 ではより高い正答率が示された。最後に、Qiu et al.(2024)は経済実験ではないが、医療保険プランの選択データ(5,998 件)を GPT-4 に模倣させ、人間のリスク下での意思決定を再現可能かを検証した。集計レベルでは人間の選好と類似した出力が得られたが、個別レベルでの一致率は低く、推定された損失回避係数(1.09)は人間の平均値(2.56)を大きく下回っており、LLM が相対的に損失を過小評価する傾向があると結論づけている。

その他、調査回答の再現性に関する先行研究としては、Santurkar et al. (2023)、Park et al. (2024)、Dominguez-Olmedo et al. (2025)などが挙げられる。これらの研究は、実際の世論調査やアンケートデータと生成 AI の出力を比較することにより、AI による回答傾向の再現性や乖離の程度を評価している。認知・心理バイアスに関する先行研究としては、Aher et al. (2023)、Binz and Schulz (2023)、Hagendorff et al. (2023)などがあり、認知的錯誤や反射的思考の再現性などに焦点を当てた検証が行われている。政治態度および投票行動に関する先行研究としては、Argyle et al. (2023)、Motoki et al. (2024)、Bisbee et al. (2024)などがあり、特定の政治的イデオロギーや属性条件の下での再現性やバイアス傾向が分析されている。さらに、その他の行動に関する先行研究としては、Brand et al. (2023)、Bybee (2023)、Lee et al. (2024)、Li et al. (2024)、Xiong et al. (2024)などがあり、市場調査、経済予測、環境意識、観光分野など多様な領域での AI と人間の類似性が検討されている。これらの詳細については、表 1 を参照されたい。

以上の通り、本節では、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を、経済実験やその他の応用領域といった多様な観点から体系的に整理・検討した。これらの研究は、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。

全体として、GPT-4 をはじめとする最新の LLM は、平均的傾向や集計レベルでの人間の判断を高精度で模倣できる能力を備えているが、極端な選択肢の再現や属性条件を考慮した場合の応答には限界がある。また、一部では「超精度バイアス」や過度な一貫性が確認されており、むしろ非人間的な出力として問題となるケースも指摘されている。さらに、社会的少数派や文化的に代表性の低い集団に関する出力では、系統的なバイアスが生じやすいことも複数の研究により報告されている。

他方で、これらの先行研究の多くは、生成 AI と人間の意思決定を比較する際に、再現性を測定するための評価指標を十分に精緻化しているとは言いがたい。とりわけ、本研究が用いるワッサースタイン距離を採用しているのは、Santurkar et al.(2023)および Xie et al.(2024)に限られている。他の研究では KL ダイバージェンスを使用して分布間の比較を試みているが、KL ダイバージェンスは分布の全体的な形状の差異を適切に捉えにくい場合があり、生成 AI と人間の回答分布の実質的な違いを捉えるには不十分である可能性がある。その意味で、ワッサースタイン距離の適用はよ

り適切な手法であると考えられる。

加えて、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、属性による条件づけを行った評価は限定的である。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。また、生成 AI に対して人間の回答をもとにフィードバックを与えることにより、再現性の改善が見込まれるものの、これを実際に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。

3. 調査設計

本研究の第一の目的は、人間および生成 AI における危険選好と参照点依存の特徴を明らかにすることにある。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 の双方についての、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、人間と生成 AI に対する調査設計と分析手法について述べる。

3.1. Web 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。202 名分の不完全な回答を除外し、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、回答者の性別や年齢、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 性格特性や時間選好、最後通牒ゲーム、トロック問題などに関する項目が含まれている。本論文では、慎重な事前分析を行った上で、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、分析の焦点を絞るために、プロスペクト理論に基づく危険選好および参照点依存のみに着目して検討を行う。表 2 では、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⁸。

<表 2: 記述統計>

⁸ 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 区分で質問を行い、それぞれの階級値として順に「300 万円」「400 万円」「600 万円」「850 万円」「1250 万円」「1500 万円」と設定した。

3.2. プロスペクト理論に関する質問

ここでは、本調査で用いた危険選好および参照点依存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参照点依存を検証する目的で、利得と損失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.

この実験では、利得と損失において、期待値が等しい A と B の 2 つの選択肢を提示し、被験者の意思決定傾向を観察している。利得に対応する PROBLEM 1 では、72%の被験者が確実に利得が得られる選択肢 A を選び、28%の被験者が確率で高額な利得が得られる選択肢 B を選択した。この結果は、多くの被験者が確実性を優先し、危険回避的な選好を示すことを示唆している。

他方、損失に対応する PROBLEM 2 では、36%が確実な損失である選択肢 A を選び、64%が確率で損失が生じる選択肢 B を選択した。多くの被験者が損失を完全に回避できる可能性に賭け、より大きな損失のリスクを受け入れたことを意味し、損失における危険愛好的な傾向を示している。

さらに注目すべき点は、PROBLEM 1 と PROBLEM 2 における選択傾向の違いである。いずれも現在の富の状態を基準とした場合、どちらの問題においても選択肢 A での金額と確率の構造が同一である。それにもかかわらず、実験の結果、両問題で選択肢を選ぶ割合が異なった。このような選好の非対称性は、利得では危険回避的だが、損失では危険愛好的という心理的傾向、すなわち参照点依存の存在を示唆している⁹。この傾向は、従来の期待効用理論が前提とする合理的意

⁹ 同様の傾向は、Kahneman and Tversky (1979) でも報告されている。同研究における設問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%と 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%と 31%であった。

PROBLEM 1.

思決定とは異なる行動パターンを浮き彫りにしており、人間の判断が参照点のような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)の問題設定に依拠したうえで、以下のような形で設問を提示した。

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。

3.3. 生成 AI への調査

本研究の中核的な課題の一つは、実際の Web 調査に基づく属性情報をもとに構成した「ペルソナ」を生成 AI に実装し、人間と同一条件下での意思決定傾向を比較可能にした点にある。具体的には、Web 調査で収集した人間の性別・年齢・世帯年収などの属性データをもとに代表的なパターンを抽出し、それをプロンプト形式で AI に入力することで、特定属性を有する仮想的な人格＝ペルソナを生成 AI に付与した。その上で、Web 調査と同一の選択課題を提示し、各ペルソナからの応答を取得することで、条件統制された比較分析が可能となった。

本研究で用いた生成 AI は、OpenAI 社のモデルである GPT-4o である。このモデルは、一般的な ChatGPT のインターフェースとは異なり、OpenAI の API 経由でのアクセスが可能であり、プログラムベースでの操作・制御が行えることが特徴である。API を活用することで、Python コードにより、大量かつ自動化された応答取得が実現可能となり、多数のペルソナ条件下での系統的な分析が

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).

可能となった¹⁰。

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、以下の手順に従って構成された。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 ファイルを参照する Python スクリプトを作成し、OpenAI 社の GPT-4o モデルと連携させて応答を生成する処理を自動化した(図 1)。

スクリプトの冒頭(1-5 行目)では、CSV ファイルの入出力処理および OpenAI API との通信に必要なライブラリ(pandas, csv, openai 等)を読み込んでいる。7 行目では、API キー(認証情報)を設定することで、OpenAI API の利用が可能となる。9-10 行目では、分析対象となる CSV ファイルのパスを指定している。12-24 行目では、生成された応答から「理由」と「回答」を抽出する関数 `extract_reason_and_answer` を定義しており、理由が明示されない場合には「不明」と出力される仕様となっている。続く 26-53 行目の `generate_reason_and_answer` 関数では、各ペルソナの属性情報(性別・年齢・年収)をもとにプロンプトを構成し、API を通じて生成 AI に応答を生成させる処理が実行される。55 行目では、DataFrame 上で上記の処理を全サンプル(N=4,838)に対して適用し、得られた理由・回答をそれぞれの変数に格納している。以降のコードにおいて、これらの応答データを CSV ファイルとして保存し、後の分析に用いている。

<図 1:生成 AI に対する質問提示・応答取得時の 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、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¹¹。この temperature の設定により、AI の選択傾向に与える確率的変動の影響を評価することが可能となる。

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 GPT-4o 上に再現し、2 つの選択課題に対する応答を取得した(図 2)。得られた応答データに対しては、次節で述べるロジック・モデルを適用し、生成 AI の危険選好および参照点依存に関する傾向を定量的に分析した。

<図 2:OpenAI API による応答例>

¹⁰ 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 7 章で記述している。

¹¹ 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。ただし、temperature が 1 よりも大きい時、分析結果が安定しない傾向が観察された。

3.4. バイアスの推定手法

本項では、回答者¹²の属性が危険選好および参照点依存に与える影響を明らかにするための推定手法について説明する。設問では、回答者は、「選択肢 1=100%の確率で 1 万円」「選択肢 2=50%の確率で 2 万円」のいずれか一方を選択する形式となっている。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、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・モデルによって推定を行うことが適切である。ロジット・モデルの選択確率は以下のように与えられる。

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

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。

4. 人間と学習前 AI¹³のバイアス比較

本節では、前章で説明した個人属性にもとづく「ペルソナ」ごとのロジット・モデルによる推定結果について説明し、人間を対象とした Web 調査より得られた回答と生成 AI の応答に基づく選択傾向の全体像を概観する。

4.1. 選択傾向の概要

表 3 は、人間を対象とした Web 調査の回答者および生成 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、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。

¹² これ以降、Web 調査の調査対象者、およびペルソナを実装した生成 AI 双方のことを「回答者」と呼ぶ。「個人」「回答」などの表現も同様である。

¹³ 4 章で扱う生成 AI の応答は、5 章以降で導入するファインチューニングを施していないものである。そのため、この生成 AI を便宜的に「学習前 AI」と呼ぶ。

<表 3:人間・学習前 AI 選択結果>

まず、人間の Web 調査の結果を確認する。利得の設問である質問 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、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、利得における危険回避的傾向を示唆している。他方、損失の設問である質問 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%、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、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的性が示された。

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 9 割近い高い選択率を示しており、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 では、生成 AI において危険愛好的性が明確に観察されるのに対し、Web 調査における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。

また、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、生成 AI ではそれぞれ 91.1% および 11.0%であり、損失においてより強い参照点依存がうかがえる。他方、Web 調査ではそれぞれ 88.4%および 57.7%であり、生成 AI と比較して相対的に弱い参照点依存がみられる。

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。ここでは $temperature$ の設定を 1.0 と、0.5、0.0 として評価を行っている。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、利得での危険回避性が観察された。他方、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的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 件のみであった。このデータは他のデータの傾向と大きく異なっており、以降の分析では当該データを除外して分析を行う。

4.2. 限界効果による分析

表 4 および表 5 は、各 $temperature$ での生成 AI および人間の Web 調査のデータを用いて、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。他方、予測確率の平均値は、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 を選ぶ確率を算出し、その平均値を求めたものである。この値は表 3 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。

<表 4: 人間・学習前 AI ロジット・モデル推定結果(利得:3 属性)>

<表 5: 人間・学習前 AI ロジット・モデル推定結果(損失:3 属性)>

まず、人間を対象とした Web 調査の推定結果を概観する。利得の質問 1 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 歳高くなるに伴い選択確率が 1.5%増加、年収が 100 万円増加すると選択確率が 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、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 歳高くなるに伴い選択確率が 2.1%増加し、年収が 100 万円増加すると選択確率が 0.8%低下するという結果が得られた¹⁴。

次に、`temperature=1.0` に設定した場合の生成 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 歳高くなると選択確率が 5.7%増加、年収が 100 万円増加すると選択確率が 2.7%低下する傾向が示された。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 歳高くなることで選択確率が 2.3%増加し、年収が 100 万円増加すると選択確率が 0.6%低下す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`)の応答結果の比較を行う。比較にあたって、デルタ法を用いて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・年収に対する限界効果の差の 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。他方で、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・年収に対する限界効果の差の z 値はそれぞれ-2.04、-0.31、-0.71 で、利得の場合とは異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がたいことが示された。

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。

¹⁴ 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。

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=1.0、0.5、0.0 でそれぞれ 0.057、0.073、0.120)。他方、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 間で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)に対して、生成 AI では temperature=1.0、0.5 でそれぞれ 0.023、0.008 であった。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 はやや大きい限界効果を示したが、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)。

4.3. 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図示したものである。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(年齢=44.9 歳、年収=593 万円)に固定したうえで、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、いずれの点においても 95%信頼区間が重なっており、人間を対象とした 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。

<図 3: 人間・学習前 AI 選択確率(利得:3 属性)>

<図 4: 人間・学習前 AI 選択確率(損失:3 属性)>

他方、「利得:3 属性 年齢」および「利得:3 属性 年収」では、人間の Web 調査の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 の線はより急峻であり、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では、95%信頼区間が重なっており、人間と各生成 AI の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 軸における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 前後に位置しており、低い年齢での選択確率に違いがみられる。対して、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ことから、年齢の増加による追加的な選択確率の変化は見られない。

さらに、両図の y 座標の水準を比較することで、参照点依存の傾向を把握することができる。たと

例えば、図3「利得:3属性 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 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。この差は、前節で述べたとおり、生成 AI における強い参照点依存を反映していると考えられる。さらに、この図から読み取れる重要な点として、人間のデータでは参照点依存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ことが挙げられる。

4.4. 的中率の分析

表6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 $\hat{P}(Y_i = 1 | \mathbf{X})$ が0.5以上であれば選択肢1を取るものとし、0.5未満であれば選択肢2をとるように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。また、ここで「的中率」とはモデルによる予測値と実際の観測値の結果が整合的な場合の割合を示している。ここでは予測値と観測値の結果がともに選択肢1である場合(1,1)の割合と、予測値と観測値の結果がともに選択肢2である場合(2,2)の割合の合計を「的中率」として定義している。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で、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調査よりも高い一致度を示している。

<表6: 人間・学習前 AI ロジット・モデルの的中率>

5. 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 に学習させることで、人間のバイアスの傾向を生成 AI で再現すること、そして学習前後の生成 AI と人間の近接距離を測ることにある。

まず、我々は生成 AI に対する学習手法として、ファインチューニングを用いた。ファインチューニングとは、大規模データで事前に訓練された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を指す(Devlin et al., 2019)。本研究で使用している OpenAI 社の GPT においては、ファイン訓練用データとして用いるためのデータセットを収集し、JSONL 形式で OpenAI の API システムにアップロードすることでファインチューニングが実行される。

続いて、ファインチューニングの実施方法について説明する。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 $\hat{P}(Y_i = 1 | \mathbf{X})$ を算出したものである。そして、学習後の生成 AI に対しても、人間や学習前 AI と同様の質問を行い、その回答をもとにロジット・モデルによ

る予測確率の推定を行うことで、三者のバイアス傾向の差異を明らかにする。

図 5 は訓練用データとして用いた JSONL 形式ファイルの抜粋である。GPT のファインチューニングに際しては、“system” “user” “assistant” の 3 つの役割(“role”)による一行の会話を 10 行以上羅列した JSONL ファイルを用意することが求められる。ここで、3 つの役割はそれぞれ、モデルが従うべき指示、使用者がモデルに投げかける質問、モデルによる回答にあたる。我々は“system”に仮想的な個人属性(ペルソナ)を、“user”に人間の回答者の選択確率を尋ねる質問を入力し、“assistant”に人間の回答者の選択確率の数値を入力した。

ファインチューニングを行うにあたっては、学習方法とパラメータの設定をあらかじめ定めておく必要がある。まず、人間の選択確率を学習させる方法として、①人間の選択確率の数値を直接入力する方法(確率による方法)と、②質問文と回答データを人間の選択頻度を反映させて入力する方法(頻度による方法)¹⁵の 2 つが考えられる。次に、ファインチューニングにおける代表的なパラメータとして、“epoch”と“Learning Rate Multiplier”(以下、LR)が挙げられる。epoch は学習の繰り返し回数を示し、値が大きいほど学習量が増える一方で、過学習のリスクも高まる。LR は学習率の乗数であり、学習の進行速度を調整するものである。値が大きいほど学習が速く効率的に進み、収束しやすくなるが、同様に過学習のリスクも高くなる。本研究では、適切な学習方法とパラメータを選定するため、あらかじめ 100 サンプルを用いて試行的にファインチューニングを実施し、学習後のモデルに対して質問への応答生成を指示したうえで、ロジット分析による評価を行った。API システムのデフォルト設定 (epoch=3, LR=2) では、質問 1 においてほぼ全てのサンプルが選択肢 1 を選ぶという偏った出力が得られた。これは過学習が生じていることを示している。そのため本研究では、epoch や LR の値をデフォルト値よりも低く設定して上記の試行を複数回行い結果を比較したうえで、最終的に学習方法として確率による方法を採用し、パラメータ設定は epoch=1, LR=0.2 を採用した。

<図 5:ファインチューニング用データセット>

図 6 は学習の際に生成 AI(GPT-4o-2024-08-06)に読み込ませた Python スクリプトである。1-3 行目の "import" によるコードは、JSONL ファイルの入出力や OpenAI API との接続に必要なモジュ

¹⁵ 後者の方法の具体例は以下の通りである。例えば年収 300 万円の 20 歳女性の質問 2 についての予測確率が 0.60 であった場合、

```
{ "messages": [ { "role": "system", "content": "あなたは 20 歳女性で、年収 300 万円です。"}, { "role": "user", "content": "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？ 1=確実に 10,000 円失う 2=50%の確率で 20,000 円失い、50%の確率で何も失わない"}, { "role": "assistant", "content": "1" } ] }
```

という会話を 60 行、会話の最後の部分を{"role": "assistant", "content": "2"}に改変した会話を 40 行の計 100 行作成する。この作業を 4,838 人の回答者、および 2 つの質問文について行い、 $100 \times 4,838 \times 2 = 967,600$ 行の JSONL データを作成する。

ールを読み込むものである。6 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 行目の "file_path" は、読み込む対象ファイルのパスを指定する。12-18 行目は、学習に用いるための JSONL ファイル(図 5)を生成 AI に読み込ませるためのものである。21-31 行目は、GPT-4o-2024-08-06 モデルに epoch=1、LR=0.2 でファインチューニングを実行するものである。33 行目から 36 行目は、ファインチューニングの進捗状況を出力させるための指示である。

<図 6: ファインチューニング時の Python スクリプト>

以上の手順によりファインチューニングを行った学習後の生成 AI に対しても、3 章 3 節で記述した学習前 AI と同様の作業を施し、Web 調査と同一の属性(性別・年齢・世帯年収)を持つ 4,838 件のペルソナを生成 AI 上に再現し、2 つの質問に対する応答を取得した。得られた応答に対して、3 章 4 節で述べたロジック・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。

6. 人間と学習前 AI・学習後 AI のバイアス比較

本節では、個人属性にもとづく「ペルソナ」ごとのロジック・モデルによる推定結果について、人間と学習前 AI の結果、およびファインチューニングにより得られた学習後 AI の結果を比較検討する。なお、生成 AI の応答結果は、いずれもデフォルト値である temperature=1.0 としたものである。

6.1. 選択傾向の概要

表 7 では、Web 調査の回答者および学習前後の生成 AI による各質問への選択結果を示している。表中の選択結果は、確実な利得または損失を示す選択肢 1 の平均選択率を表す。まず、学習後 AI の結果を確認する。利得の質問 1 では、確実に 1 万円が得られる選択肢 1 を選択した割合は 96.1% であり、選択肢 2 を選択した割合(3.9%)を大きく上回った。この結果は、利得における高い危険回避的傾向を示唆している。他方、損失の質問 2 では、確実に 1 万円を失う選択肢 1 を選択した割合は 40.4% であり、選択肢 2 を選択した割合は 59.6% であった。多くの応答が確率的に損失が生じる選択肢 2 を選択しており、損失における危険愛好性が示されたが、利得における危険回避性ほど強い傾向とは言えない。

<表 7: 人間・学習前後 AI 選択結果>

続いて、人間と学習前後の AI の結果を比較する。利得の質問 1 では、人間と学習前 AI、学習後 AI が選択肢 1 を選択した割合はそれぞれ 88.4%、91.1%、96.1% であった。学習後 AI は、人間よりも平均選択率が高い学習前 AI と比べても、さらに高い選択率がみられた。このことより、利得における危険回避性の過学習が示唆される。同様に損失の質問 2 では、人間と学習前 AI、学習

後 AI が選択肢 1 を選択した割合はそれぞれ 57.7%、11.0%、40.4%であった。生成 AI の平均選択率は、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。

また、質問 1 および質問 2 における選択肢 1 の選択率を比較すると、学習後 AI では 96.1%および 40.4%であり、学習前 AI(91.1%および 11.0%)と比較して参照点依存が弱く、人間の損失回避傾向(88.4%および 57.7%)に近づいている。

6.2. 限界効果の分析

表 8 および表 9 は、学習前後の生成 AI および Web 調査のデータを用いて、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。予測確率の平均値は表 7 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。

<表 8:人間・学習前後 AI ロジット・モデル推定結果(利得:3 属性)>

<表 9:人間・学習前後 AI ロジット・モデル推定結果(損失:3 属性)>

まず、利得の質問 1 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 1.9%高くなり、年齢が 10 歳高くなることで選択確率が 3.0%増加し、年収が 100 万円増加することで選択確率が 1.4%低下する傾向が見られた。他方、損失の質問 2 では、性別および年齢においてのみ統計的に有意な限界効果が観察され、年収による影響は確認されなかった。具体的には、女性である場合に選択肢 1 を選ぶ確率が 26.4%高くなり、年齢が 10 歳高くなることで選択確率が 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。

次に、人間と学習後 AI 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、年収の限界効果の差についてのz値は、それぞれ 2.14、-4.16、6.36 であり、いずれの属性においても、統計的に有意な差が確認された。同様に質問 2 では、性別、年齢、年収のz値がそれぞれ-15.2、-17.5、-1.77 となり、性別と年齢については有意な差が認められたが、年収については有意な差は見られなかった。

続いて、学習前 AI と学習後 AI の応答結果を比較する。質問 1 における性別、年齢、年収のz値は、それぞれ 1.63、7.49、-9.19 であり、性別についてのみ有意な差が見られず、年齢および年収については有意な差が認められた。同様に質問 2 では、性別、年齢、年収のz値がそれぞれ-16.6、-19.4、-1.06 であり、性別・年齢については有意な差が認められたが、年収については有意

な差は見られなかった。

6.3. 選択確率の図解

図 7 および図 8 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。図 7 から読み取れる傾向として、まず「利得: 3 属性 性別」では、いずれの点においても、学習前 AI と人間の 95%信頼区間が重なっており、両者の選択確率の差は確認できない。しかし、学習後 AI は学習前 AI と人間に比べて上方に位置しており、学習により危険回避度が高くなったことが分かる。

<図 7: 人間・学習前後 AI 選択確率(利得:3 属性)>

<図 8: 人間・学習前後 AI 選択確率(損失:3 属性)>

他方、「利得: 3 属性 年齢」および「利得: 3 属性 年収」では、生成 AI の折れ線の傾きが、学習を通じて緩やかになり、相対的に人間の折れ線の傾きに近づいている。ただし、いずれのグラフにおいても平均値付近では、学習前 AI と人間の 95%信頼区間が重なっているのに対し、学習後 AI はそれらよりも上方に位置しており、この点からも、学習による利得の危険回避度の高まりが読み取れる。

図 8 からは全く別の傾向が読み取れる。「損失: 3 属性 性別」および「損失: 3 属性 年齢」では、学習後 AI は、いずれも高い正の限界効果を反映し傾きが急峻である。そのため、性別が男性である場合や、年齢が低い場合は学習前 AI の折れ線に近づき、性別が女性である場合や、年齢が高い場合は人間の折れ線に近づく。また、両者とも平均値付近では、学習後 AI の y 座標が、人間と学習前 AI の y 座標の中間に位置しており、損失の危険愛着性を適切に学習できていることが見てとれる。「損失: 3 属性 年収」では、3 本の折れ線の傾きには大きな違いが見られないが、y 切片には明確な差が認められる。学習後 AI の折れ線が、他 2 本の中間かつ人間の折れ線寄りに位置していることが分かる。

さらに、属性の平均値周りで両図の y 座標の水準を比較することで、参照点依存の傾向を視覚的に把握することができる。この図から読み取れる重要な点は、学習により生成 AI における顕著な損失回避傾向が弱まり、生成 AI が人間の損失回避傾向に近づいたという点である。

6.4. 的中率の分析

表 10 は、各モデルにおける選択確率の予測精度を示すものであり、実際の人間による選択結果との一致度(的中率)を評価している。学習後 AI の応答に基づく的中率は、質問 1 において 96.57%と非常に高く、人間(88.40%)および学習前 AI(92.08%)を上回る。これは、学習後 AI が利

得における人間の意思決定傾向を高い精度で再現していることを示唆している。

他方、質問 2(損失)における学習後 AI の的中率は 69.93%であり、人間(57.59%)よりは高いものの、学習前 AI(88.98%)よりも低い水準にとどまった。これは、学習後 AI が参照点依存に関する人間の行動を再現するうえで一定の限界があることを示している。

<表 10: 人間・学習前後 AI ロジット・モデルの的中率>

6.5. ワッサースタイン距離による選択確率の比較分析

本項では、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。特に、ユークリッド距離の 2 乗をコスト関数とする場合、2 次ワッサースタイン距離 $W_2(\mu^1, \mu^2)$ は以下のように定式化される (Ambrosio et al., 2008):

$$W_2(\mu^1, \mu^2) := \min \left\{ \int_{\mathcal{X}^2} d(x_1, x_2)^2 d\mu(x_1, x_2) : \mu \in \mathcal{T}(\mu^1, \mu^2) \right\}$$

ここで μ^1, μ^2 は確率空間上の任意の確率分布であり、 $\mathcal{T}(\mu^1, \mu^2)$ は μ^1 と μ^2 を周辺分布に持つ結合確率分布の集合を、 $d(x_1, x_2)$ は二点 x_1, x_2 間のユークリッド距離を表している。式の右辺は分布 μ^1 から分布 μ^2 への輸送計画の中で、コスト関数 $d(x_1, x_2)^2$ に対する最小の輸送コストを表している。この距離指標を用いることで、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 11 には、人間と学習前 AI、および学習後 AI の選択確率分布間における 2 次ワッサースタイン距離を示している。

<表 11: 三者間の二次ワッサースタイン距離>

まず、利得の質問 1 において、人間と学習前 AI の距離は 0.135 であり、人間と学習後 AI の距離は 0.123 であった。また、学習前 AI と学習後 AI の距離は 0.081 であった。このことから、人間と AI の距離は 0.012(約 8.9%) 短縮されたことが分かる。さらに、学習後 AI は人間と比べて 52%(=1-0.123/0.081)だけ学習前 AI 寄りであることが分かる。

次に、損失の質問 2 では、選択確率分布の距離はそれぞれ、0.467、0.263、0.353 であった。これにより、学習によって人間と AI の距離は 0.204、すなわち約 43.7%縮まったと言える。さらに、学習後 AI は学習前 AI と比べて 34%(=1-0.263/0.353)だけ人間寄りであることが分かる。

最後に、質問 1・2 を合算したデータ(N=9,676)に基づき、その二次ワッサースタイン距離を測る

と、選択確率分布の距離はそれぞれ、0.343、0.198、0.253 であった。このことから、学習により生成 AI と人間との距離が 42.3%縮まり、学習後 AI は学習前 AI と比べて $28\%(=1-0.198/0.253)$ だけ人間寄りとなったと言える。

以上の結果から、総じて言えば、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。

7. 考察

7.1. 人間と学習前 AI のバイアス傾向について

4 章の分析結果から、生成 AI(GPT-4o)は利得において、人間と同様に危険回避的な選好傾向を示しており、選択行動の再現性という点で高い整合性が確認された。特に、質問 1 における選択肢 1 の選択率は、人間・AI ともに 9 割前後に達しており、プロスペクト理論に基づく行動パターンを忠実に反映していると評価できる。他方で、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 AI の間に大きな差は見られなかったものの、年齢および年収に関しては、いずれも生成 AI の方が限界効果の絶対値が大きくなる傾向が確認された。また、生成 AI は損失においても危険回避的な選択を行っており、人間と比較してより明確な選好の非対称性を示している。

生成 AI は訓練データに含まれる属性と選好のパターンを過剰に再現しており、いわば「バイアスの過学習(overfitting)」が生じている可能性がある。すなわち、実際の人間が示すばらつきや非一貫性を再現するのではなく、文脈上尤もらしいと判断される応答を選び続けることで、属性ごとの典型的傾向を過度に強調するように出力が生成されていると考えられる。とりわけ、**temperature** を 0.0 に設定した条件では、他の設定と比べて係数の絶対値が顕著に大きくなっている。出力のランダム性が抑制されることによって、人間の判断に見られるような不確実性や曖昧さが失われ、より機械的かつ規範的なパターンが反映されやすくなっていることが分かる。このように、生成 AI の応答は文脈依存的かつ属性主導的な出力傾向を内包しており、それがバイアス構造の再現において人間との重要な相違点を生じさせていると考えられる。

他方、損失においては、生成 AI は人間よりも顕著な参照点依存を示した。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、より強い参照点依存を示した。さらに、**temperature**=0.0 の条件では、1 つのサンプルを除いて、すべてのサンプルが選択肢 2 のみを選択しており、損失を極端に回避する傾向が観察された。

また、参照点依存に関しても、生成 AI において同様の傾向が観察された。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%であり、今回の Web 調査では 88%対 58%であったものの、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 となっていた。このことから、人間と AI で危険選好や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。

GPT-4o 以外の大規模言語モデル(LLM)—すなわち、OpenAI 社の GPT-3.5、Google 社の Gemini 2.0 Flash、DeepSeek 社の DeepSeek-R1—に対して同様の設問を提示した結果、各モデル間に明確な応答傾向の違いが認められた。GPT-3.5 については、危険選好に関する理解が不十分であり、Gemini 2.0 Flash においては設問の意図を的確に把握していないと見受けられる応答が散見された。他方、DeepSeek-R1 は、危険選好や参照点依存に関する既存研究の知見を高い精度で反映し、GPT-4o と類似した応答傾向を示していた。以上より、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危険選好や損失回避に関する応答の整合性が向上し、一部の行動経済学的判断傾向がより忠実に反映されるようになった可能性がある。また、Gemini における当該特性の学習は現時点では限定的にとどまっており、DeepSeek は GPT-4o と同等水準で当該特性を獲得している可能性が示唆される。

7.2. 学習後 AI のバイアス傾向について

分析結果から、生成 AI(GPT-4o)は利得において、人間が持つ危険回避的な選好傾向を過学習することが確認された。質問 1 における選択肢 1 の選択率は、人間が 88%で、学習前 AI が 91%、学習後 AI が 96%であった。そのため、プロスペクト理論に基づく危険回避性を反映してはいるものの、学習によりその傾向が強まっていることを示している。この結果は、前節で述べた「バイアスの過学習(overfitting)」が生じていることを示唆している。他方、属性ごとの限界効果を比較すると、性別については学習により人間と AI との間の差が広がったものの、年齢および年収に関しては、いずれも人間と AI との差が縮まる傾向にあることが確認された。ただし、3 属性とも人間と学習後 AI の限界効果の間には有意差がみられた。2 次ワッサースタイン距離を用いて評価を行うと、学習によって生成 AI の選択傾向が 1 割近く人間に近付いていることが認められる。

次に、損失においては、学習により AI の危険回避性が弱まり、より人間に近付くことが確認された。学習前には生成 AI は質問 2 における選択肢 1 の選択率が 11.0%と強い危険回避性を示していたが、学習により Tversky and Kahneman(1988)の結果(36%)よりも大きな選択率(40.4%)を記録し、今回の Web 調査の結果(57.7%)に大きく近づいた。2 次ワッサースタイン距離を見ても、人間と生成 AI 間の近接距離は 4 割以上縮まっており、損失においては利得よりも学習が進んでいることが分かる。さらに、表 11 で示した的中率の比較においても、学習前 AI におけるすべてのサンプルが予測値 2 となる傾向はなくなり、学習により 65%以上のサンプルが予測値 1 をとることが確認された。しかし、属性ごとの限界効果に着目すれば、学習により人間と AI 間の差が広まった。特に性別、年齢においては、学習後 AI はそれぞれ性別で 0.264、年齢で 0.133 と高い限界効果を示し、属性が危険回避性に与える影響については過学習が示唆される。

さらに、参照点依存に関しても、学習により AI の参照点依存が弱まり、人間に近付いたことが観察された。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、学習前 AI ではおおむね 90% 対 10% となっているのに対し、学習後 AI では 96% 対 40% となった。Tversky and Kahneman(1988)では 72% 対 36% であり、本研究における人間の回答では 88% 対 58% であることから、学習により AI の参照点依存が弱まったことが分かった。

8. 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。そのうえで、人間の意思決定傾向を生成 AI に学習させることで、人間の傾向を生成 AI で再現することを目指し、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。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、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、学習によって生成 AI の選択傾向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。特に損失においての学習では、人間と AI 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、この点は学習に際して改善の余地があるといえる。

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、他のバイアスに関し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. Aher, G. V., Arriaga, R. I., & Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. *In International Conference on Machine Learning* (pp. 337-371). PMLR.
2. Akata, E., Schulz, L., Coda-Forno, J., Oh, S. J., Bethge, M., & Schulz, E. (2023). Playing repeated games with large language models. *arXiv preprint arXiv:2305.16867*.
3. Ambrosio, L., Gigli, N., & Savaré, G. (2008). Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media.
4. Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3), 337-351.
5. Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120.
6. Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4), 401-416.
7. Brand, J., Israeli, A., & Ngwe, D. (2023). Using GPT for market research. *Harvard business school marketing unit working paper*, (23-062).
8. Brookins, P., & DeBacker, J. (2024). Playing games with GPT: What can we learn about a large language model from canonical strategic games?. *Economics Bulletin*, 44(1), 25-37.
9. Brown, A. L., Imai, T., Vieider, F. M., & Camerer, C. F. (2024). Meta-analysis of empirical estimates of loss aversion. *Journal of Economic Literature*, 62(2), 485-516.
10. Bybee, L. (2023). Surveying Generative AI's Economic Expectations. *arXiv preprint arXiv:2305.02823*.
11. 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*.
12. Demszky, D., Yang, D., Yeager, D.S., Bryan, C.J., Clapper, M., Chandhok, S., Eichstaedt, J.C., Hecht, C., Jamieson, J., Johnson, M., Jones, M., Krettek-Cobb, D., Lai, L., JonesMitchell, N., Ong, D.C., Dweck, C.S., Gross, J.J., & Pennebaker, J.W. (2023). Using large language models in psychology. *Nature Reviews Psychology*, 2(11), 688-701.
13. Dominguez-Olmedo, R., Hardt, M., & Mendler-Dünner, C. (2025). Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37, 45850-45878.
14. 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.
15. Guo, F. (2023). GPT in game theory experiments. *arXiv preprint arXiv:2305.05516*.
16. Hagendorff, T. (2023). Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988*, 1.
17. 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.
18. Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? (No. w31122). National Bureau of Economic Research.
19. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1, pp. 4171-4186.

20. 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.
21. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 363-391.
22. Lee, S., Peng, T. Q., Goldberg, M. H., Rosenthal, S. A., Kotcher, J. E., Maibach, E. W., & Leiserowitz, A. (2024). Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Climate*, 3(8), e0000429.
23. Li, P., Castelo, N., Katona, Z., & Sarvary, M. (2024). Frontiers: Determining the validity of large language models for automated perceptual analysis. *Marketing Science*, 43(2), 254-266.
24. Lorè, N., & Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1), 18490.
25. Macmillan-Scott, O., & Musolesi, M. (2024). (Ir) rationality and cognitive biases in large language models. *Royal Society Open Science*, 11(6), 240255.
26. Mei, Q., Xie, Y., Yuan, W., & Jackson, M. O. (2024). A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9), e2313925121.
27. Motoki, F., Pinho Neto, V., & Rodrigues, V. (2024). More human than human: measuring ChatGPT political bias. *Public Choice*, 198(1), 3-23.
28. Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P., Bernstein, M.S. (2024). Generative agent simulations of 1,000 people. *arXiv preprint*, arXiv:2411.10109.
29. Qiu, L., Singh, P. V., & Srinivasan, K. (2023). Consumer Risk Preferences Elicitation From Large Language Models. Available at SSRN 4526072.
30. Rahwan (2018). Society-in-the-Loop: Programming the Algorithmic Social Contract. *Ethics and Information Technology*, 20, 5–14.
31. Rahwan et al. (2019). Machine Behaviour. *Nature*, 568, 477–486.
32. Ross, J., Kim, Y., & Lo, A. W. (2024). Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint*, arXiv:2408.02784.
33. 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.
34. Sarstedt, M., Adler, S. J., Rau, L., & Schmitt, B. (2024). Using large language models to generate silicon samples in consumer and marketing research: Challenges, opportunities, and guidelines. *Psychology & Marketing*, 41(6), 1254-1270.
35. Tsuchihashi, T. (2023). How much do you bid? Answers from ChatGPT in first-price and second-price auctions. *Journal of Digital Life*, 3.
36. 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.
37. Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5, 297-323.
38. Xie, Y., Liu, Y., Ma, Z., Shi, L., Wang, X., Yuan, W., ... & Mei, Q. (2024). How Different AI Chatbots Behave? Benchmarking Large Language Models in Behavioral Economics Games. *arXiv preprint* arXiv:2412.12362.
39. Xiong, X., Wong, I. A., Huang, G. I., & Peng, Y. (2024). Understanding AI-generated experiments in tourism: replications using GPT simulations. *Journal of Travel Research*, 00472875241275945.

図表

表 1:生成 AI と人間の意思決定・バイアス比較を行っている先行研究

論文番号	著者	調査対象バイアス・実験	使用 AI	対象・サンプル数	平均的再現性	条件付き再現性	人間からの FB	メトリクス
経済学的な意思決定とバイアス								
1	Akata et al. (2023)	経済実験 (Prisoner's Dilemma, Battle of the Sex)	GPT-3, 3.5, 4	LLM のみ:ペイオフ行列を変えて各 10 ラウンド、合計 1,224 試行実施	該当なし (LLM の結果のみ) 囚人のジレンマ・ゲームでは高パフォーマンス、男女の争いゲームではパフォーマンス低い	該当なし	該当なし	該当なし
2	Guo (2023)	経済実験 (最後通牒ゲーム、囚人のジレンマ)	GPT-4 (gpt4-1106-preview)	LLM のみ (先行研究と比較): 利己的・公平性の特性を付与 (UG400, PD300)	先行研究の結果と類似	該当なし	プロンプト調整のみ	該当なし
3	Horton (2023)	経済実験 (社会選好: Charness and Rabin (2002), 公正さ: Kahneman et al. (1986), 現状維持バイアス: Samuelson and Zeckhauser (1988), 最低賃金: Horton(2023))	GPT-3 (davinci-003 他)	先行研究の結果の再現	既存の結果と高い類似度	該当なし	RLHF あり	該当なし
4	Tsuchihashi (2023)	経済実験 (オークションでの入札行動)	GPT-3.5	LLM のみ (先行研究と比較): FPA, SPA 各 40 ラウンド	FPA で先行研究と類似した傾向 (FPA で過剰入札、SPA でやや正確)	該当なし	プロンプト調整のみ	該当なし
5	Brookins & DeBacker (2024)	経済実験 (独裁者ゲーム、囚人のジレンマ)	GPT-3.5-turbo	LLM のみ (先行研究と比較): Dictator: 500 回, PD: 1,100 回	LLM は人間よりも公平・協力的な回答傾向	リスク・誘惑・効率性で評価。効率性指標のみで類似した傾向	プロンプト調整のみ	該当なし
6	Jia et al. (2024)	経済実験 (損失回避, 確率加重, リスク選好)	ChatGPT-4.0-Turbo, Claude-3-Opus, Gemini-1.0-pro	LLM のみ (先行研究と比較)	ある程度人間の選択と類似した傾向	属性(性別等)でバイアスが変動(人間の解答との比較無し)	プロンプト調整のみ	該当なし
7	Lorè & Heydari (2024)	経済実験 (2 人ゲームにおける戦略行動)	GPT-3.5, GPT-4, LLaMa-2	LLM のみ: 4 ゲーム×5 文脈×3 モデル×300 回	該当なし (LLM の結果のみ) GPT-4 で構造依存、GPT-3.5 は文脈依存	該当なし	プロンプト調整のみ	該当なし
8	Mei et al. (2024)	経済実験 (利他性, 公平性, 信頼, 協力的性)	GPT-3.5-Turbo, GPT-4	公開データ約 9 万人と比較, LLM データ: 6 ゲーム×30 回	ChatGPT-4 で人間と類似した結果	該当なし	プロンプト調整のみ	該当なし
9	Qiu et al. (2024)	保険プランの選択	GPT-4-turbo	欧米の 2006 年までの保険データ (5,998 件)	集計レベルでは人間の選択と類似した選択傾向を示す 個人ベースではあてはまりが悪い	LLM の回答結果から心理パラメータを推定: 損失回避係数・確率重み係数は先行研究より小さな値を取る傾向	該当なし	F1 スコア
10	Ross et al. (2024)	経済実験 (不平等回避, 損失回避, 時間割引)	GPT-3.5, GPT-4, Claude 2 他 9 種	LLM のみ (先行研究と比較)	人間と異なる傾向: 不平等回避: 他者への罪悪感強いが、嫉妬は弱い 損失回避: 人間よりも利益に対しては合理的だが、損失に対しては非合理的 時間割引: 人間よりも強い割引傾向	該当なし	プロンプト調整のみ	該当なし
11	Xie et al. (2024)	経済実験 (利他性, 公平性, リスク, 協力的性)	GPT-4o, LLaMa3, Claude 3, 他	LLM のみ (先行研究と比較): 6 ゲーム×5 モデル×50 回	LLM の分布は人間と類似した傾向 公平な選択を行い、高い協力率がみられる	該当なし	プロンプト調整のみ	ワッサースタイン距離
12	Chen et al. (2025)	18 種の認知バイアス	GPT-3.5-turbo, GPT-4	LLM のみ (バイアスの有無のみを判定)	多くのバイアスが再現される。 GPT-3.5 では損失が回避的で、利得でリスク愛好的 GPT-4 はフレーミングに関係なくリスク回避的傾向	該当なし	該当なし	該当なし
13	Macmillan-Scott & Musolesi (2024)	12 の認知課題	GPT-3.5, GPT-4, Claude2, Bard, LLaMA	LLM のみ (既存の結果との比較)	人間らしいバイアスの割合が高いモデルは GPT-3.5。 最も人間らしい回答傾向がみられるのは GPT-4。	該当なし	該当なし	該当なし

アンケート・調査データの再現性								
14	Santurkar et al. (2023)	世論調査データ	GPT モデル, AI21 Labs モデル	米国 (ATP 世論調査 2017~2021 年)	LM の意見は、一般的な米国民の意見と大きなズレがある 特定の政治的傾向 (リベラル寄り) を示す傾向がみられる	該当なし	RLHF あり	ワッサースタイン距離
15	Park et al. (2024)	性格診断・行動実験 (経済・心理実験 15 種)	LLM を活用したエージェント・アーキテクチャ	1,052 人の AI 代理シミュレーション	参加者の 2 週間後の回答を高い精度で再現 (85%程度)	年齢・人種・政治イデオロギーで条件付けて評価。 タスク間のバイアスを一貫して減少させる傾向。	プロンプト調整のみ	該当なし
16	Dominguez-Olmedo et al. (2025)	アンケートデータの再現 (順序・ラベルバイアス)	GPT-2~GPT-4, Llama 他 43 モデル	2019 年の ACS 25 問	全体として非常に低い	該当なし	RLHF あり	KL ダイバージェンス
認知・心理バイアス								
17	Aher et al. (2023)	チューリング実験 (Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds)	GPT-3.5, GPT-4 (ほか 8 種)	LLM のみ (先行研究と比較): 各タスク 1,000 仮想被験者、数千応答	LM-5 以降で人間に近い再現度 超精度バイアスがみられる場合もある (群衆の知恵タスク)	該当なし	RLHF あり	該当なし
18	Binz & Schulz (2023)	心理学実験 (意思決定能力、情報探索能力、熟慮能力、因果推論能力)	GPT-3 (Davinci)	LLM のみ (先行研究と比較): 数千問反復 (13,000 等)	誤答パターン含め一部人間に類似	該当なし	プロンプト調整のみ	該当なし
19	Hagendorff et al. (2023)	心理学実験 (CRT・semantic illusion)	GPT-1~GPT-4	人間: 455 人、LLM: 各タスク×50 回×モデル 10 種	GPT-3 までは直観的な回答傾向 GPT-4 では人間よりも高い正答率 (正答率 96%)	該当なし	プロンプト調整のみ	該当なし
政治的態度と投票行動								
20	Argyle et al. (2023)	政治的態度・投票行動	GPT-3	ANES 1,304 人 + 2,873 人	投票行動で相関 (>0.9)	一部の 카테고리 (例: 無党派層) で再現性低	プロンプト調整のみ	Cramér's V, テトラコリック相関
21	Bisbee et al. (2024)	政治的ペルソナ設定による ANES データ比較	ChatGPT 3.5 Turbo, ChatGPT 4.0, Falcon-40B	米国 (ANES2016・2020 年)	LLM の回答は人間の回答と類似性高いがばらつき少ない	政治的属性で見た場合、人間と LLM の回答は、ばらつきは異なるが同様の傾向を示す 個人属性で条件つけた場合、人間の回答と LLM の回答は異なる傾向を示す	RLHF あり	該当なし
22	Motoki et al. (2024)	政治的バイアス	GPT-3.5	LLM のみ: 民主党・共和党になりすまして回答・非なりすまし回答 (各 100 回答)	人間との比較無し デフォルトの回答は民主党よりの回答 設定したペルソナによって政治的立場が変化する	該当なし	プロンプト調整のみ	該当なし
その他								
23	Brand et al. (2023)	複数製品に対する WTP	GPT-3.5-turbo-0125	LLM のみ (先行研究と比較)	平均的な WTP パターンは再現 新製品については人間の回答との整合性が低下する傾向	一部の製品カテゴリーについて、人間と類似した傾向がみられる	ファインチューニングあり	該当なし
24	Bybee (2023)	経済期待予測 (WSJ 記事ベース)	GPT-3.5	米国 (1984-2021 の WSJ 記事 300 件)	SPF, AAIL, CFO などの既存調査と高相関	該当なし	該当なし	該当なし
25	Lee et al. (2024)	温暖化に関する信念	GPT-3.5-turbo-16k, GPT-4	米国全国調査 (2017 年: 1,304, 2021 年: 1,006)	温暖化の信念については高い再現度 (85%) 原因や懸念については再現度低い (51%/48%)	黒人層の意見を過小評価	プロンプト調整のみ	F1 スコア, Cramér's V
26	Li et al. (2024)	ブランド認知・知覚的類似	GPT-4, GPT-Neo	21 自動車ブランド、530 人の評価	高い一致率 (87.2%)	年齢・属性間で一貫性あり	プロンプト調整のみ	Triplet 一致率
27	Xiong et al. (2024)	感情・信念形成 (観光領域)	GPT-3.5-turbo	LLM のみ (先行研究と比較): 16 シナリオ×100 回	LLM の回答結果は人間の回答結果と類似傾向	該当なし	プロンプト調整のみ	該当なし

表 2: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。

```

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:***¥***.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？
29     1. 確実に 10,000 円失う
30     2. 50%の確率で 20,000 円失い、50%の確率で何も失わない
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=1.0
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:***¥***.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}")

```

図 1: 生成 AI に対する質問提示・応答取得時の Python スクリプト

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。また、4 章におけるファインチューニング後の質問提示・応答取得の際には、40 行目の“gpt-4o”をファインチューニング済モデル名に置き換えた。

	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

図 2: OpenAI API による応答例

Notes: 3 章 3 節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)を示している。実際には結果は、csv 形式で出力される。

表 3:人間・学習前 AI 選択結果

質問 1 利得	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 損失	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

Notes:表中の「AI」は全て「学習前 AI」のことを指している。

表 4:人間・学習前 AI ロジット・モデル推定結果(利得:3 属性)

質問 1 利得		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
McFadden R^2		0.6355	0.3932	0.3472	0.0140
予測確率		0.898 (0.004)	0.889 (0.005)	0.911 (0.004)	0.884 (0.005)

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。ここでは、表中の「AI」は全て「学習前 AI」のことを指している。

表 5:人間・学習前 AI ロジット・モデル推定結果(損失:3 属性)

質問 2 損失		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
McFadden R^2		-	0.0644	0.0204	0.0038
予測確率		-	0.030 (0.002)	0.110 (0.004)	0.577 (0.007)

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。ここでは、表中の「AI」は全て「学習前 AI」のことを指している。

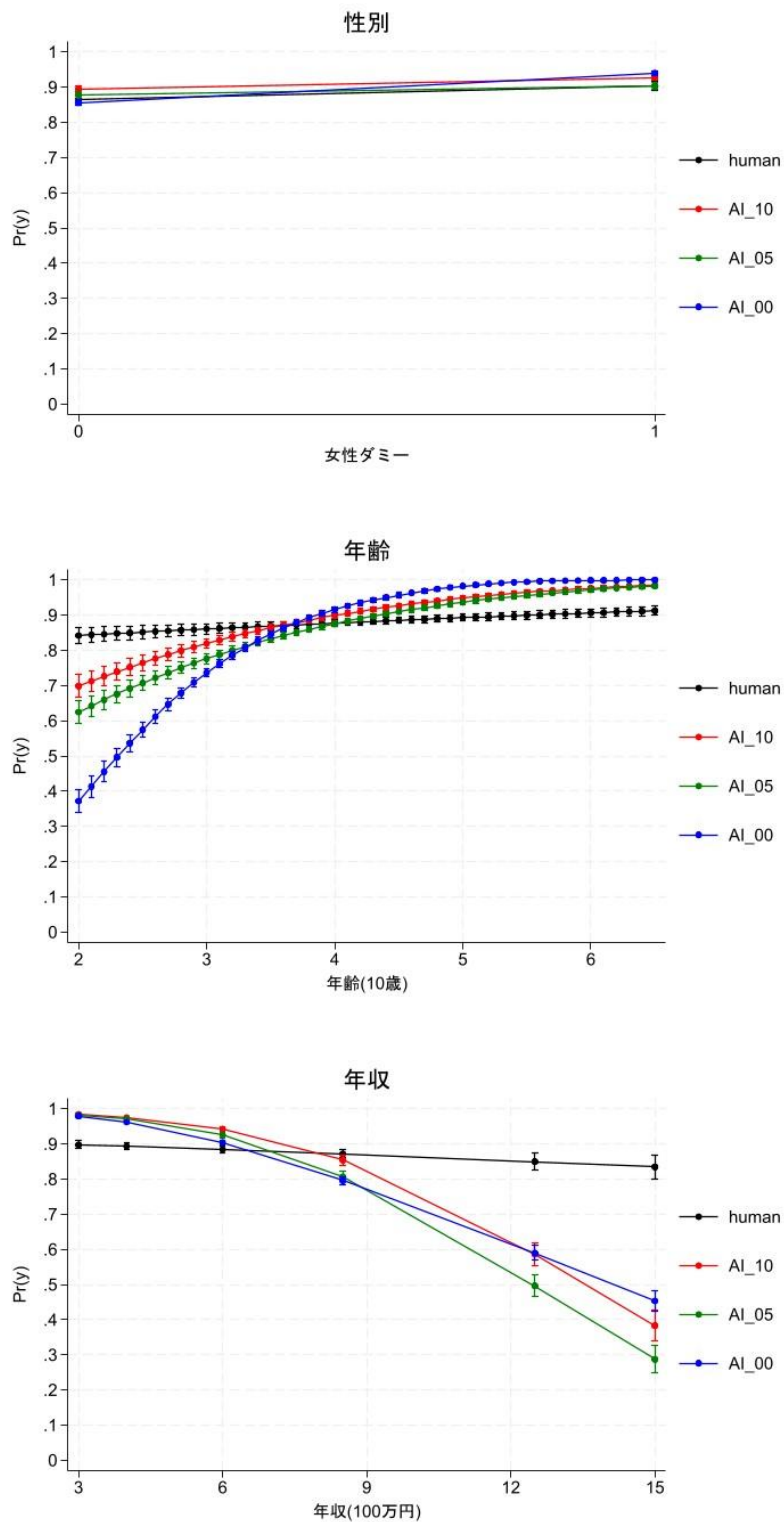


図 3:人間・学習前 AI 選択確率(利得:3 属性)

Notes: AI_10 は、学習前 AI に temperature 1.0 で尋ねた場合のことを指している。AI_05、AI_00 も同様である。また、図中の各点における上下の棒は、95%信頼区間を表す。

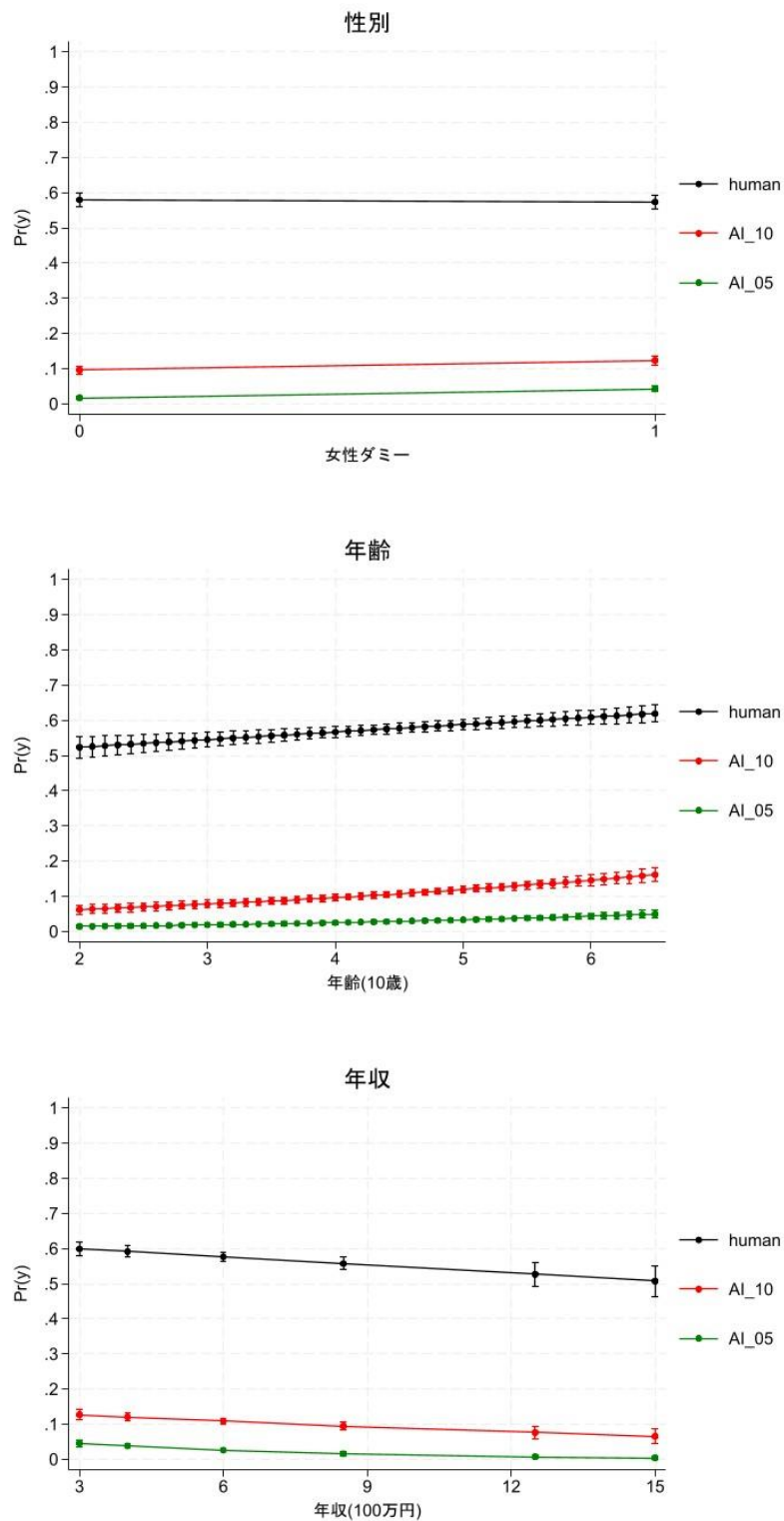


図 4: 人間・学習前 AI 選択確率(損失:3 属性)

Notes: AI_10 とは、学習前 AI に 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、図中の各点における上下の棒は、95%信頼区間を表す。

表 6: 人間・学習前 AI ロジット・モデルの的中率

質問 1 利得	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 損失	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%

Notes: 表中の「AI」は全て「学習前 AI」のことを指している。

```

{"messages": [{"role": "system", "content": "あなたは 36 歳男性で、年収 300 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.58"}]}
{"messages": [{"role": "system", "content": "あなたは 57 歳女性で、年収 300 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.62"}]}
{"messages": [{"role": "system", "content": "あなたは 40 歳女性で、年収 850 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.54"}]}
{"messages": [{"role": "system", "content": "あなたは 36 歳男性で、年収 300 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.87"}]}
{"messages": [{"role": "system", "content": "あなたは 57 歳女性で、年収 300 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.93"}]}
{"messages": [{"role": "system", "content": "あなたは 40 歳女性で、年収 850 万円です。"}, {"role": "user", "content": "質問「あなたは現在の富に上乘せして 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」の選択肢 1 の選択確率は？"}, {"role": "assistant", "content": "0.89"}]}

```

図 5: ファインチューニング用データセット

Notes: ファインチューニングの際に用いた JSONL ファイルの抜粋。各質問における選択確率をもとに訓練を行った。

```

1 import openai
2 import os
3 import json
4
5 # ◆ OpenAI APIキーを設定
6 openai.api_key = os.getenv("OPENAI_API_KEY")
7
8 # ◆ アップロードする JSONL ファイルのパス
9 file_path = r"C:\Users\***\***.jsonl"
10
11 # ◆ ファイルを OpenAI にアップロード
12 print("Uploading training file...")
13 response = openai.files.create(
14     file=open(file_path, "rb"),
15     purpose="fine-tune"
16 )
17 file_id = response.id
18 print(f"File uploaded successfully! File ID: {file_id}")
19
20 # ◆ ファインチューニングの実行 (GPT-4oを使用)
21 print("Starting fine-tuning process...")
22 fine_tune_job = openai.fine_tuning.jobs.create( # 最新API仕様
23     training_file=file_id,
24     model="gpt-4o-2024-08-06",
25     hyperparameters={
26         "n_epochs": 1,
27         "learning_rate_multiplier": 0.2
28     }
29 )
30 fine_tune_id = fine_tune_job.id
31 print(f"Fine-tuning started! Job ID: {fine_tune_id}")
32
33 # ◆ ファインチューニングの進捗を確認
34 print("Checking fine-tuning status...")
35 status_response = openai.fine_tuning.jobs.retrieve(fine_tune_id) # 最新API仕様
36 print("Fine-tuning status:", status_response.status)

```

図 6: ファインチューニング時の Python スクリプト

Notes: epoch=1、LR=0.2 として、図 5 のデータセットを gpt-4o モデルに読み込ませることでファインチューニングを行った。

表 7:人間・学習前後 AI 選択結果

質問 1 利得	学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
Mean	0.911	0.961	0.884
Std. Dev.	0.285	0.193	0.320
Obs.	4,838	4,838	4,838

質問 2 損失	学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
Mean	0.110	0.404	0.577
Std. Dev.	0.313	0.491	0.494
Obs.	4,838	4,838	4,838

表 8:人間・学習前後 AI ロジット・モデル推定結果(利得:3 属性)

質問 1 利得		学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
限界効果	女性ダミー	0.033*** (0.007)	0.019*** (0.005)	0.041*** (0.009)
	年齢 (10 歳)	0.057*** (0.003)	0.030*** (0.002)	0.015*** (0.003)
	所得 (100 万円)	-0.027*** (0.001)	-0.014*** (0.001)	-0.005*** (0.001)
obs.		4,838	4,838	4,838
McFadden R^2		0.3472	0.4500	0.0140
予測確率		0.911 (0.004)	0.961 (0.003)	0.884 (0.005)

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 9:人間・学習前後 AI ロジット・モデル推定結果(損失:3 属性)

質問 2 損失		学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
限界効果	女性ダミー	0.028*** (0.009)	0.264*** (0.011)	-0.006 (0.014)
	年齢 (10 歳)	0.023*** (0.004)	0.133*** (0.004)	0.021*** (0.005)
	所得 (100 万円)	-0.006*** (0.002)	-0.003 (0.002)	-0.008*** (0.002)
obs.		4,838	4,838	4,838
McFadden R^2		0.0204	0.1804	0.0038
予測確率		0.110 (0.004)	0.404 (0.007)	0.577 (0.007)

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。

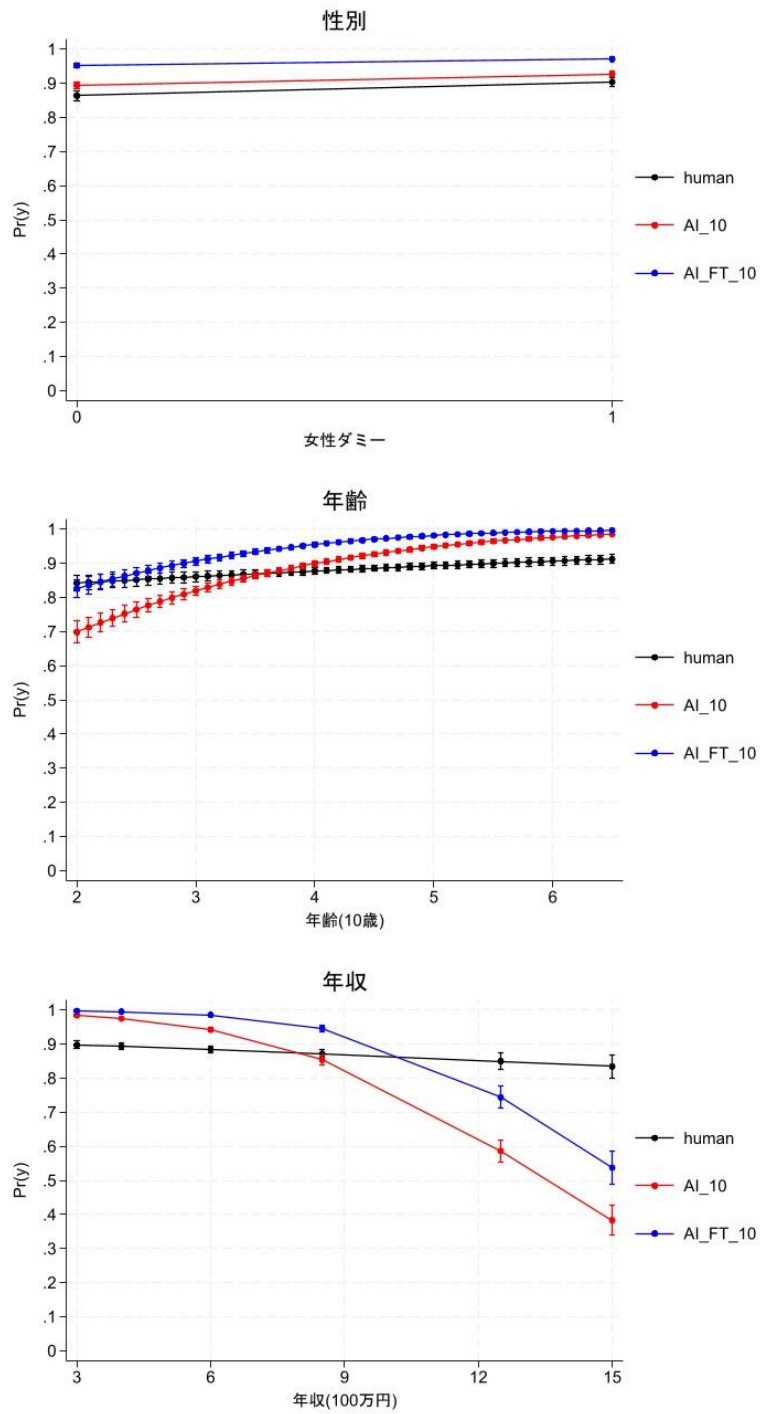


図 7:人間・学習前後 AI 選択確率(利得:3 属性)

Notes: AI_10とAI_FT_10はそれぞれ、学習前AIに temperature 1.0 で尋ねた場合と学習後AIに temperature 1.0 で尋ねた場合のことを指している。また、図中の各点における上下の棒は、95%信頼区間を表す。

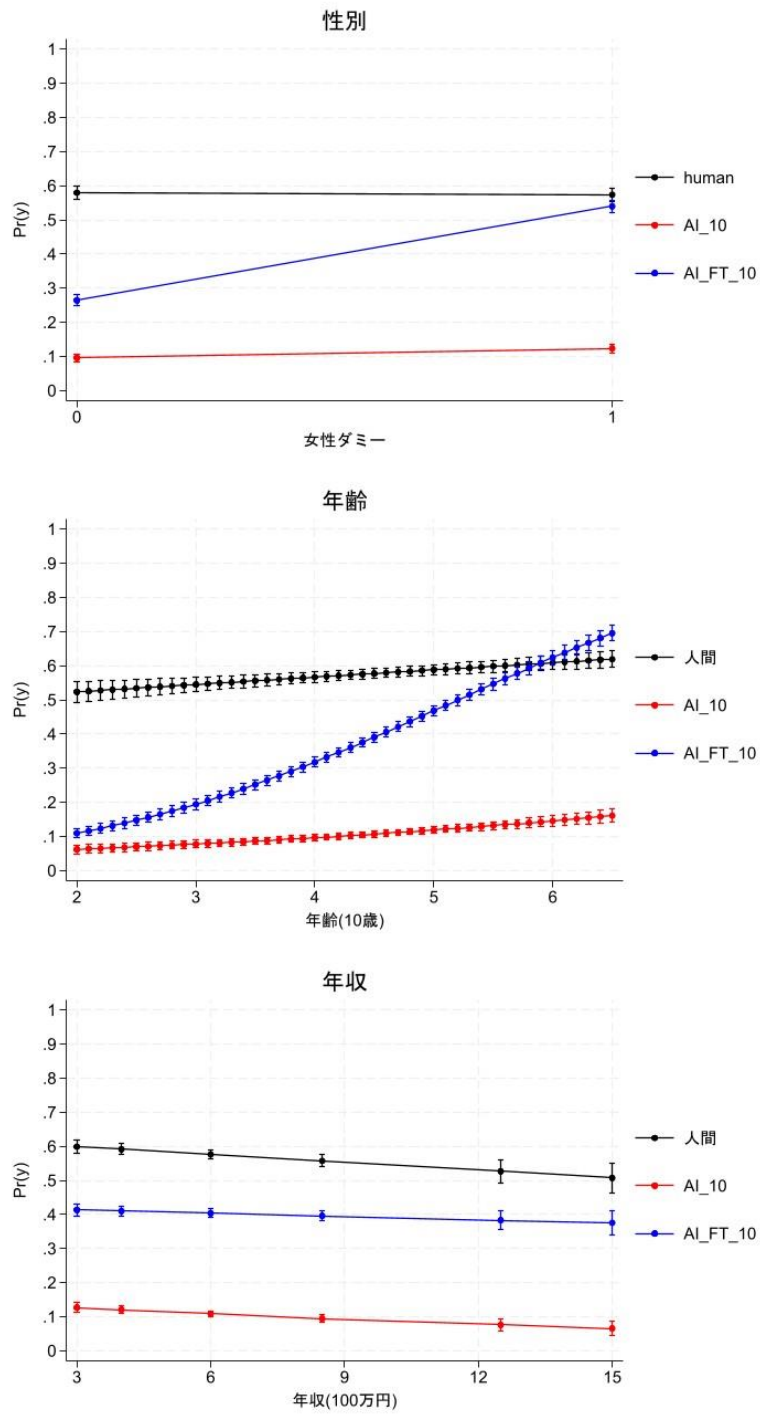


図 8:人間・学習前後 AI 選択確率(損失:3 属性)

Notes: AI_10とAI_FT_10はそれぞれ、学習前AIに temperature 1.0 で尋ねた場合と学習後AIに temperature 1.0 で尋ねた場合のことを指している。また、図中の各点における上下の棒は、95%信頼区間を表す。

表 10: 人間・学習前後 AI ロジット・モデルの的中率

質問 1 利得	学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
予測値,観測値			
(2, 2)	2.34% (n=113)	1.24% (n=60)	0% (n=0)
(2, 1)	1.34% (n=65)	0.78% (n=38)	0% (n=0)
(1, 2)	6.57% (n=318)	2.65% (n=128)	11.60% (n=561)
(1,1)	89.75% (n=4,342)	95.33% (n=4,612)	88.40% (n=4,277)
的中率	92.08%	96.57%	88.40%

質問 2 損失	学習前 AI temperature 1.0	学習後 AI temperature 1.0	人間
予測値,観測値			
(2, 2)	88.98% (n=4,305)	47.66% (n=2,306)	0.93% (n=45)
(2, 1)	11.02% (n=533)	18.13% (n=877)	1.03% (n=50)
(1, 2)	0% (n=0)	11.95% (n=578)	41.38% (n=2,002)
(1,1)	0% (n=0)	22.26% (n=1,077)	56.66% (n=2,741)
的中率	88.98%	69.93%	57.59%

Notes: 表中の百分率の数値は小数点第 3 位で四捨五入したものであるため、(2,2)と(1,1)の値の和は、必ずしも的中率と一致するわけではない。

表 7: 三者間の二次ワッサースタイン距離

	人間・ 学習前 AI 間	人間・ 学習後 AI 間	学習前 AI・ 学習後 AI 間
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

Notes: 表中の値は全て temperature=1.0 で取得した応答に基づく分析である。

Appendix

Human-in-the-Loop 機械学習と行動経済学——アノテーションと能動学習——

Human-in-the-Loop 機械学習とは、“Human-in-the-Loop”と機械学習を融合させた概念である。HITL は元来、監視制御システムの文脈で研究されてきた概念であり、人間とシステムが相互に補完しあいながら運用を改善していくループを指す(Monarch 2021; Samad 2020)。機械学習研究の進展に伴い、両者を統合した「HITL 機械学習」という概念が導入されるようになった(Horvitz et al., 1999)。近年では、AI の発展とともに、HITL 機械学習の AI 分野への応用が進んでおり、とりわけ機械学習のラベル付け作業に人間が関与する取り組みとして、AI の発展に不可欠の要素となっている¹⁶。

HITL 機械学習のプロセスは、(1)能動学習(2)アノテーション(3)機械学習モデルによる学習(4)機械学習によるラベル予測の 4 つの要素から構成される(図 A-1-a)。能動学習とは、ラベル付けに用いるデータのサンプリングプロセスを指す。代表的な手法として、ランダム・サンプリングや不確実性サンプリング¹⁷が挙げられる。アノテーションとは、データに対するラベル付け作業を指す。HITL 機械学習においては、人間が主体的に関与する能動学習およびアノテーションの戦略立案が特に重要視される。

< 図 A-1: Human-in-the-Loop 機械学習のプロセス >

Monarch (2021)は、HITL 機械学習の代表的な事例として、ニュースの見出し(短い文章)の解析、特定のトピックに関する情報の文章からの抽出、画像中の特定の物体の識別の 3 例を挙げている。一方、AI をペルソナ化し、選好傾向を分析する本研究のプロセスについても、上記の 3 例と並ぶ HITL 機械学習の実践例として位置付けられる。以下ではこの考えに基づき、HITL 機械学習のプロセスの各要素と本研究の対応関係を論じる。

本研究における能動学習の対応要素は、Web 調査により収集した人間回答者(N=4,838)のアンケートデータのうち、回答者全員分のデータから、性別・年齢・世帯所得の 3 つの個人属性、および危険選好に関する 2 つの質問の回答を抽出した点である。本研究の目的は、Web 調査で得た 4,838 人の人間回答者の選好傾向を AI 上で再現することにある。このプロセスは、AI の「ペルソナ化」および AI からの応答取得プロセスの基盤となる。

本研究におけるアノテーションの対応要素は、特定の個人属性(性別・年齢・所得)を付与し、「ペルソナ化」した生成 AI(GPT-4o)4,838 サンプルのそれぞれに対し、人間回答者から得

¹⁶ 初期の代表例として、大規模画像データベースである“ImageNet”が挙げられる(Russakovsky et al., 2015)。

¹⁷ モデルの予測が不確実、すなわちモデルが予測困難な、紛らわしいデータのみを集中的にサンプリングする手法。

た2択の質問の回答に基づく危険選好の傾向を付与した点である。本研究において具体的に「ラベル」として用いたのは、各回答者の予測確率である。これは、二値の回答を被説明変数とし、性別・年齢・世帯年収の3属性を説明変数として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 $\hat{P}(Y_i = 1 | \mathbf{X})$ を算出したものである。

著者らが二値回答そのものではなく、ロジット・モデルより得た予測確率をアノテーションに使用したことには、2つの利点が存在する。第一に、過学習(overfitting)を防げる点である。ロジット・モデルの構造を介することにより、人間回答者の選好傾向を0以上1以下の連続値で表すことができる。これにより、個人属性に紐づけられた選好傾向を、1または2という二値によって固定し、強調することを緩和できる。第二に、人間・AI間のバイアス測定・再学習・近接度測定に関する一貫した分析が可能となる点である。本研究では、AIへの再学習用データとして、人間のロジット・モデル選択確率を用い、さらにAIの人間への接近度合の評価指標として、ロジット・モデル選択確率に基づいて導出したワッサースタイン距離を使用した。以上のように、ロジット・モデル選択確率をアノテーションに使用したことで、一貫通貫した定量的メトリクスに基づく分析が可能となった。

本研究では、機械学習モデルを学習させるプロセスに、ファインチューニングを採用した。ファインチューニングは、転移学習の一手法として、HITL機械学習でも一般的に使用される。人間とAIのロジット・モデル選択確率に差があるため、人間の選択確率をAIに学習させることで、AIの出力改善を図る。

本研究におけるラベル予測プロセスとは、再学習を経たAIを「ペルソナ化」し、学習後AIから2択の質問の応答を得たうえで、その応答をもとにロジット・モデル選択確率を導出する過程を指す。このプロセスにより、AIの人間への近接度測定が可能となる。

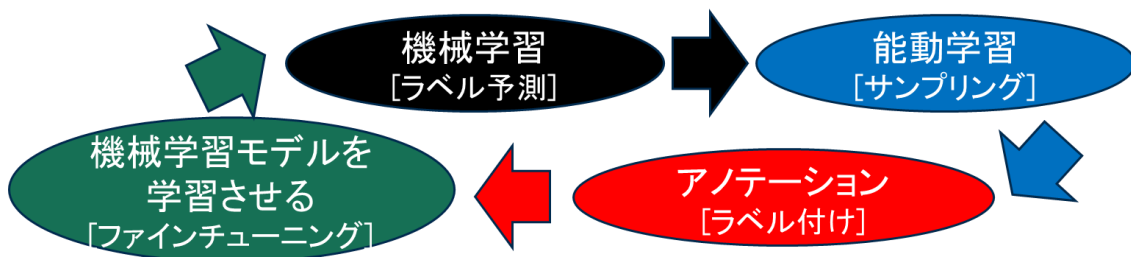
以上で述べたように、本研究のプロセス1~4からなるループは、HITL機械学習の実践となっている(図A-1-b)。プロセス3で学習を経たAIに対し、追加で他のタスクに関する人間回答者の情報(例: 最後通牒ゲームの結果や時間選好に関する質問の応答)を読み込ませることで、このループをさらに循環させることができる。

HITL機械学習における能動学習やアノテーションといった概念は、経済実験における回答者抽出・回答データ取得のプロセスに類似している。本研究は、HITL機械学習のプロセスを、一貫通貫した定量的メトリクスに基づき、行動経済学において実装した先進的事例として位置付けられる。

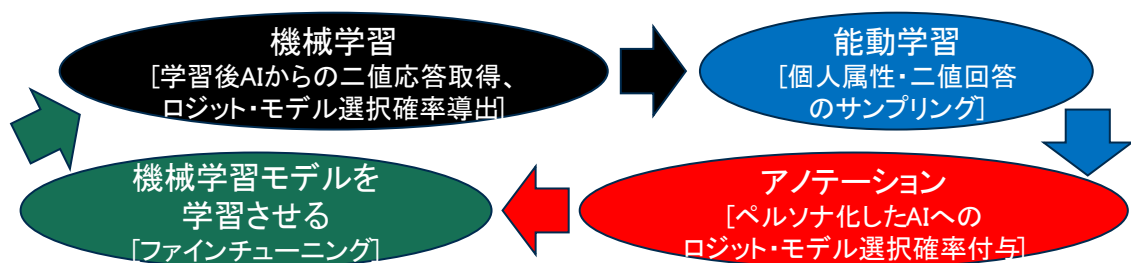
Appendix 参考文献

1. Horvitz, E. (1999). Uncertainty, action, and interaction: In pursuit of mixed-initiative computing. *IEEE Intelligent Systems*, 14(5), 17-20.
2. Monarch, Robert.1 (2021). *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Manning Publications.
3. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211-252.
4. Samad, T. (2020). Human-in-the-loop control and cyber–physical–human systems: applications and categorization. *Cyber–physical–human systems: fundamentals and applications*, 1-23.

図表



(a) データのラベルを予測する HITL 機械学習のプロセス



(b) 本研究における HITL 機械学習のプロセス

図 A-1: Human-in-the-Loop 機械学習のプロセス

Notes: Monarch (2021)を参考に著者作成。