2026年1月9日

行動経済学から見るAIの未来:
人間とAIのバイアスの比較分析

京都大学大学院経済学研究科　教授
依田　高典

第1部
行動経済学から見るAIの未来

「生成AIとは何か」を生成AIに説明させる

- 人工知能(AI)の一種：テキスト、画像、音声、動画などの多様なコンテンツを生成する技術。
- 基盤モデルに基づく技術：大規模なデータセットでトレーニングされた言語モデル(例: GPT、BERT)や画像モデル(例: DALL-E、Stable Diffusion)を活用。
- 主な用途：

  テキスト生成: チャットボット、記事作成、コード補完。

  画像生成: デザイン、自動イラスト作成。

  音声・動画生成: 音声合成、仮想キャラクター制作。
- 特徴：ユーザーの指示(プロンプト)を基に出力を生成。コンテンツの創造性や品質はモデルのトレーニング次第で変化。
- メリット：効率的なコンテンツ生成。新しい創作やアイデア支援。
- 課題：バイアスや倫理的問題。真偽不明の情報生成リスク。

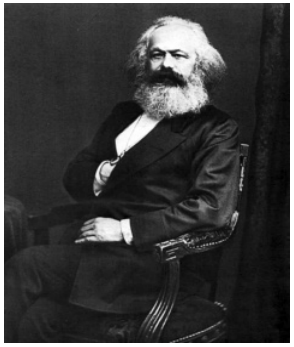
(ChatGPT-4oで作成)

AIの社会インパクト

- プロになるには１万時間のトレーニングが必要とされる。ラフに言えば、24時間×365日(=8760時間)以上かかる。
- したがって、１日１時間の訓練ならば20年以上、１日10時間の訓練ならば２年以上かかるところ、生成AIを使えば、語学・プログラミング・絵画・音楽・・・等の習熟時間を大幅に短縮できる。
- 肌感覚で言えば、既存のものを再生産するだけなら、習熟時間は10分の一以下になるのではないか。
- つまり、皆が生成AIを使うようになると、専門家の付加価値は大幅に下がり、非専門家の付加価値は大幅に上がり、付加価値の格差が小さくなる。
- ホワイトカラーを中心に、その半分が仕事を奪われるという試算には現実味がある。
- 特に、知的サポート業務については、すぐにも人間が生成AIに置き換えられる。21世紀のラッダイト運動が起きかねない。

19世紀の機械化と労働

- 19世紀のマルクスの時代、産業革命によって、資本の有機的構成の高度化が進み、労働の生産物における価値構成が低下し、失業者が増大すると予言された。
- 19世紀は人口増大の時代であり、都市への廉価な労働力の流入により、腕利きの熟練労働者の失業が生まれ、女性・子供など非熟練労働者の労働環境の劣悪化が進んだ。
- マルサスは人口増加の悲惨な未来を予測し、マルクスは労働者の大量失業を予測したので、経済学は「陰鬱な科学」と呼ばれた。
- 20世紀にケインズ経済学が登場し福祉国家が成立し、大量生産・大量消費のサイクルが出来たために、大恐慌・階級対立は何とか避けられた。



https://ja.wikipedia.org/wiki/
カール・マルクス



https://ja.wikipedia.org/wiki/
トマス・ロバート・マルサス



https://ja.wikipedia.org/wiki/
ジョン・メイナード・ケインズ

21世紀のAI化と労働

- 19世紀の機械化がブルーカラーに与えた影響と21世紀のAI化がホワイトカラーに与える影響は似ている。
- 先進国では人口減少時代に入っている。人間一人が生活するコストが増えた経済的帰結に過ぎない。このトレンドは変えられない。
- 企業としては熟練労働力を必要としているが、実際には労働力の確保に苦労している。日本のような人口減少国で、企業が直面するのは労働力不足である。
- エネルギー環境食糧問題を考えても、80億の人口は地球全体のキャパシティオーバーであり、人口減少は歓迎すべきことである。
- 労働時間が減るのであれば、余暇を楽しみ、精神的豊かさを大切にしながら、物質的豊かさを維持できることが必要である。
- この数十年、一部の経営幹部がフローの生産物価値以上の分配を占有するために、富の集中による貧富の格差が拡大している。
- 生成AIの有効活用で、少ない労働時間でも生産物価値を維持して、エネルギー環境食糧問題の解決をはかることが大切である。

行動経済学とはどういう学問か

- ホモエコノミカス(経済人)批判
- サイモン(１９７８ノーベル賞)・カーネマン(２００２ノーベル賞)・セイラー(２０１７ノーベル賞)
- 限定合理性・ヒューリスティクス・バイアス・ナッジ
- バイアスには生理学的・遺伝学的基礎がある(分かっていても止められない)。
- 前文明時代には、理性に頼らずに、感情による刹那的な生き方がむしろ合理的だった。
- 産業革命以降、寿命が延び、感情に沿った刹那的生き方が悪となった。
- 現代人の生き辛さの根源であり、進化心理学では「エデンの園追放仮説」と呼ばれる。



https://ja.wikipedia.org/wiki/
ハーバート・サイモン



https://ja.wikipedia.org/wiki/
ダニエル・カーネマン



https://ja.wikipedia.org/wiki/
リチャード・セイラー

行動経済学と人工知能

- 人工知能がニューラルネットワークを模範とする以上、行動経済学と深く関係する。
- サイモンは、第一世代人工知能学者のリーダーであるアレン・ニューウェルを高く評価し共同研究を行った。
- 人間の意思決定を支援するコンピュータシステムの開発に携わり、彼らの研究成果は1956年ダートマス会議で発表された。
- 2000年代の第三世代人工知能ブームは深層学習に端を発するが、経済学では画像データ・音声データの利用がんでいない。
- 他方で、因果推論では経済学に一日の長があり(2019年・2021年ノーベル賞)、経済学と因果的機械学習との融合が進む。
- 人間の目から見て極めて自然なテキスト・画像・音声などを生成する生成AIが登場(OpenAIの提供するChatGPTなど)。
- 生成AIの経済学の利用はこれからであるが、B2BからB2Cへの汎用的活用が期待される。

AIのはらむ社会問題

- 生成AIは企業側にも消費者側にも社会的リスクが指摘されている。
- 生成AIを生産に使う側：知財侵害・情報漏洩・評判リスク
- 生成AIを消費に使う側：プライバシー・セキュリティ・ハルシネーション
- 生成AIは、対話相手が人間か機械かを判別する「チューリングテスト」を突破し、「不気味の谷」も乗り越えつつある。
- 人間の自然知能でハルシネーションを見極めることはできない。
- 人類学者山極壽一の群れの最適規模。10～15人ならば、阿吽の呼吸で通じる。100～150人ならば、言語で意思疎通できる。
- デジタル・サイバー空間では、100～150人という物理的時空限界を超えて、接点を持つ人数が爆発的に増える。
- 匿名性の限界もあり、言語を通じて「共感」が得られにくい。
- 炎上しやすいＳＮＳ空間に、AI人格を活用して、サイバー空間の意思疎通の質を高められないか。

AIとバイアス

- バイアスは生理学的な根深いところに起因するので、気をつけても矯正することは容易でなく、矯正には大きな精神的苦痛を伴う。
- 行動経済学のナッジを提唱したサンスティーンは、「個別化されたデフォルト」を設定することを提案している。
- 生成AIの自動化と個別最適化を活かして、一人一人に最適な「AIパートナー」が求められる。
- タバコのような嗜癖の予防、弱者に対する差別的言動の改善、省エネのような社会的行動の促進に対して、社会的規範を押しつけるのではなく、生活改善を褒賞するナッジを考える。
- 学習データは汚れきったバイアスの固まりであるから、入力時と出力時にクリーニングするとはいえ、AI自体もバイアスを持つ。
- 人間のバイアスとAIのバイアスが相互作用して、社会の不安定性が増幅するような最悪のシナリオも考えられる。
- 人間の意思決定のバイアスを研究する行動経済学と生成AIの融合研究のポテンシャルは大きい。

AIナッジとヒューマン・イン・ザ・ループ

- 「ナッジ」とは、セイラーらが提唱した行動経済学の概念で、人間のバイアスを前提にしながら、提供する情報や選択肢の配置を工夫し、自発的により良い選択を促す手法を指す。
- 「AIナッジ」とは、リバタリアン・パターナリズムに則り、AIがパーソナルに最適化されたナッジを設計・提示する。
- 「Human in the loop (HITL)」とは、人間の判断や介入を、システムの一部として組み込む設計手法を指す。

フェーズ	AIナッジの役割	HITLの関与
設計前	対象の行動分析	人間がAIに学習させる変数や行動データを選定
実装	ナッジ提示の自動化	人間が「AIナッジ」の文言やタイミングをレビュー
運用	効果の測定・検証	人間がモニタリングして「バイアス」が生じないか確認
改善	学習による最適化	人間が最終判断し、社会的妥当性のフィードバックを与える

AIと行動経済学の近未来

- 生成AIの自動化と個別最適化が一層進化すると、人間側は自然知能だけで抵抗するのは難しい。
- 悪質業者の詐欺も巧妙化するので、消費者が自然知能でサービスの真贋を見抜くことも難しい。
- 「AIパートナー」が擬人的であった方が良いのか、どういった性格を持つべきか、どの程度のバイアスまで許容するべきなのか、答えはまだ不確定で、仮に答えがあるにしても、個別にしかありえない。
- 認知判断力の弱い若年者や高齢者でも、デジタル社会のコミュニケーションをサポートしてくれるような仕組みが必要である。
- プラットフォーマー等に対して、消費者保護の観点から「AIパートナー」の開発を要請し、予防的に善処してもらうようにしたい。
- デジタル社会の「ゲートキーパー」として社会的責任を果たすように努力義務を課すのも一案である。

第II部
AIと人間のバイアスの比較分析

【事例】人間とAIの損失回避性バイアス

- 第一問　あなたは、2つの選択肢のうち、どちらを選ぶか。
  ①100%確実に、100ドルをもらう。
  ②50%の確率で、200ドルをもらう。
  →　人間「選択肢①を選ぶ」。GPT-4「選択肢①を選ぶ」。
- 第二問　あなたは、2つの選択肢のうち、どちらを選ぶか。
  ①100%確実に、100ドルを失う。
  ②50%の確率で、200ドルを失う。
  →　人間「選択肢②を選ぶ」。GPT-4「選択肢②を選ぶ」。
- 人間は損失回避性のために、選好の逆転が起きる。
- AIもまた損失回避性のために、選好の逆転が起きる。
  「私自身の場合、リスク回避型かリスク許容度の高いかによって答えが変わりますが、一般的なリスク回避の傾向を考慮すると、私は確実に100ドルを得る選択肢Aを選びます。」
- 以下、損失回避性を取り上げ、詳細に分析する。

損失回避性のペルソナ分析

- 人間5千人を対象にしたWEB調査。
  利得・損失に分けた損失回避性問題の選択データ。
- 人間WEB調査データをロジット回帰して選択確率を出力。
  被説明変数：確実選択肢、説明変数：性別・年齢・所得。
- ChatGPTから5千個の選択出力をとる。
  人間と同じ「ペルソナ」を持つ損失回避性問題の選択データ。
  AIデータをロジット回帰して選択確率を出力。
- 人間とAIのロジット回帰の推定結果を比較。
  人間とAIの損失回避性、限界効果、正解率を比較。





【利得】確実選択肢を選ぶ比率の比較

- 利得の確実性バイアスに対して、属性のペルソナを設定
- 性別：男女均等、年齢：平均45歳、世帯所得：平均593万円
- 生成AI：確実選択肢①を選ぶ割合91%(N=4,838)
- 人間：確実選択肢①を選ぶ割合88%(N=4,838)

生成AI		
平均確率	0.911	***
	(0.004)	
女性	0.033	***
	(0.007)	
年齢	0.057	***
(10歳)	(0.003)	
世帯所得	-0.027	***
(100万円)	(0.001)	
McFadden R^2	0.347	

人間		
平均確率	0.884	***
	(0.005)	
女性	0.041	***
	(0.009)	
年齢	0.015	***
(10歳)	(0.003)	
世帯所得	-0.005	***
(100万円)	(0.001)	
McFadden R^2	0.014	

【利得】属性の回帰係数の比較

- ロジット・モデルの回帰分析(限界効果)
- 確実性効果は女性が大きい(AI：3%・人間：4%)
- 確実性効果は年齢が高いほど大きい(AI：6%・人間：2%)
- 確実性効果は所得が高いほど小さい(AI：-3%・人間：-1%)

生成AI		
平均確率	0.911	***
	(0.004)	
女性	0.033	***
	(0.007)	
年齢	0.057	***
(10歳)	(0.003)	
世帯所得	-0.027	***
(100万円)	(0.001)	
McFadden R^2	0.347	

人間		
平均確率	0.884	***
	(0.005)	
女性	0.041	***
	(0.009)	
年齢	0.015	***
(10歳)	(0.003)	
世帯所得	-0.005	***
(100万円)	(0.001)	
McFadden R^2	0.0140	

【損失】確実選択肢を選ぶ比率の比較

- AIの確実性バイアスに対して、ペルソナを設定
- 性別：男女均等、年齢：平均45歳、世帯所得：平均593万円
- 生成AI：確実選択肢①を選ぶ割合11%・・・損失回避性が強い
- 人間：確実選択肢①を選ぶ割合58% ・・・損失回避性が弱い

生成AI		
平均確率	0.110	***
	(0.004)	
女性	0.028	***
	(0.009)	
年齢	0.023	***
(10歳)	(0.004)	
世帯所得	-0.006	***
(100万円)	(0.002)	
McFadden R^2	0.020	

人間		
平均確率	0.577	***
	(0.007)	
女性	-0.006	
	(0.014)	
年齢	0.021	***
(10歳)	(0.005)	
世帯所得	-0.008	***
(100万円)	(0.002)	
McFadden R^2	0.004	

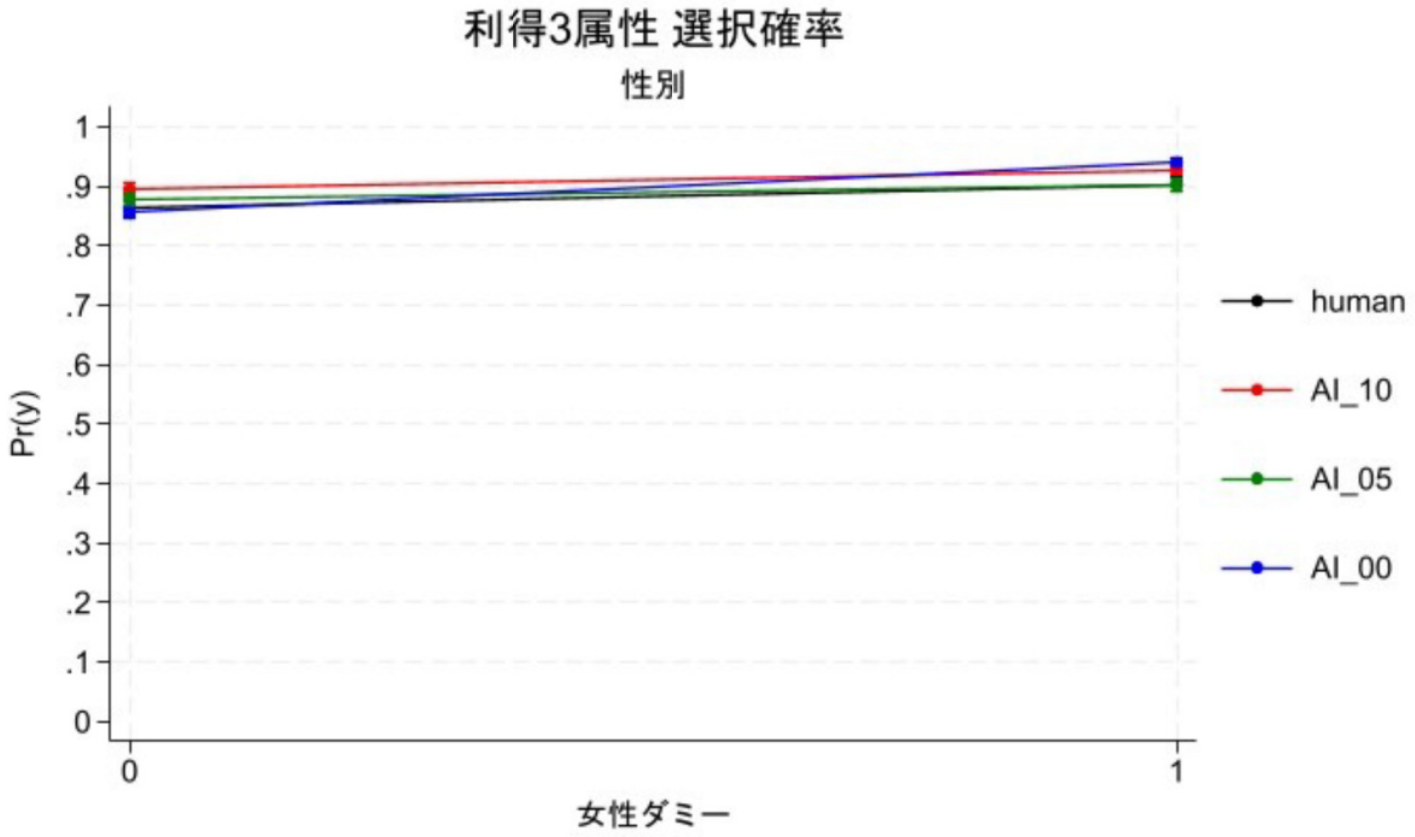
【損失】属性の回帰係数の比較

- ロジット・モデルの回帰分析(限界効果)
- 確実性効果は女性が大きい(AI：3%・人間：-1%)
- 確実性効果は年齢が高いほど大きい(AI：2%・人間：2%)
- 確実性効果は所得が高いほど小さい(AI：-1%・人間：-1%)

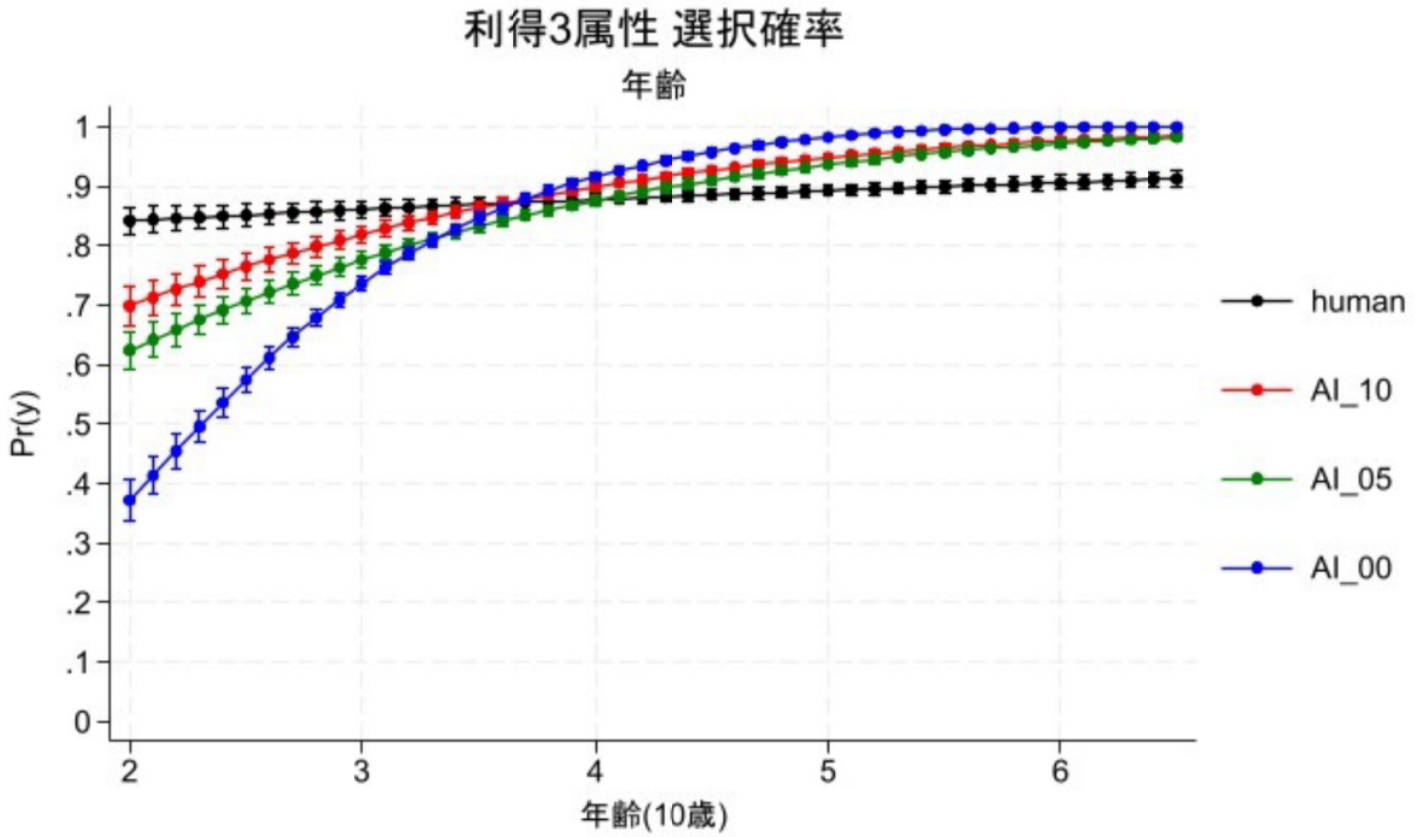
生成AI		
平均確率	0.110	***
	(0.004)	
女性	0.028	***
	(0.009)	
年齢	0.023	***
(10歳)	(0.004)	
世帯所得	-0.006	***
(100万円)	(0.002)	
McFadden R^2	0.020	

人間		
平均確率	0.577	***
	(0.007)	
女性	-0.006	
	(0.014)	
年齢	0.021	***
(10歳)	(0.005)	
世帯所得	-0.008	***
(100万円)	(0.002)	
McFadden R^2	0.0038	

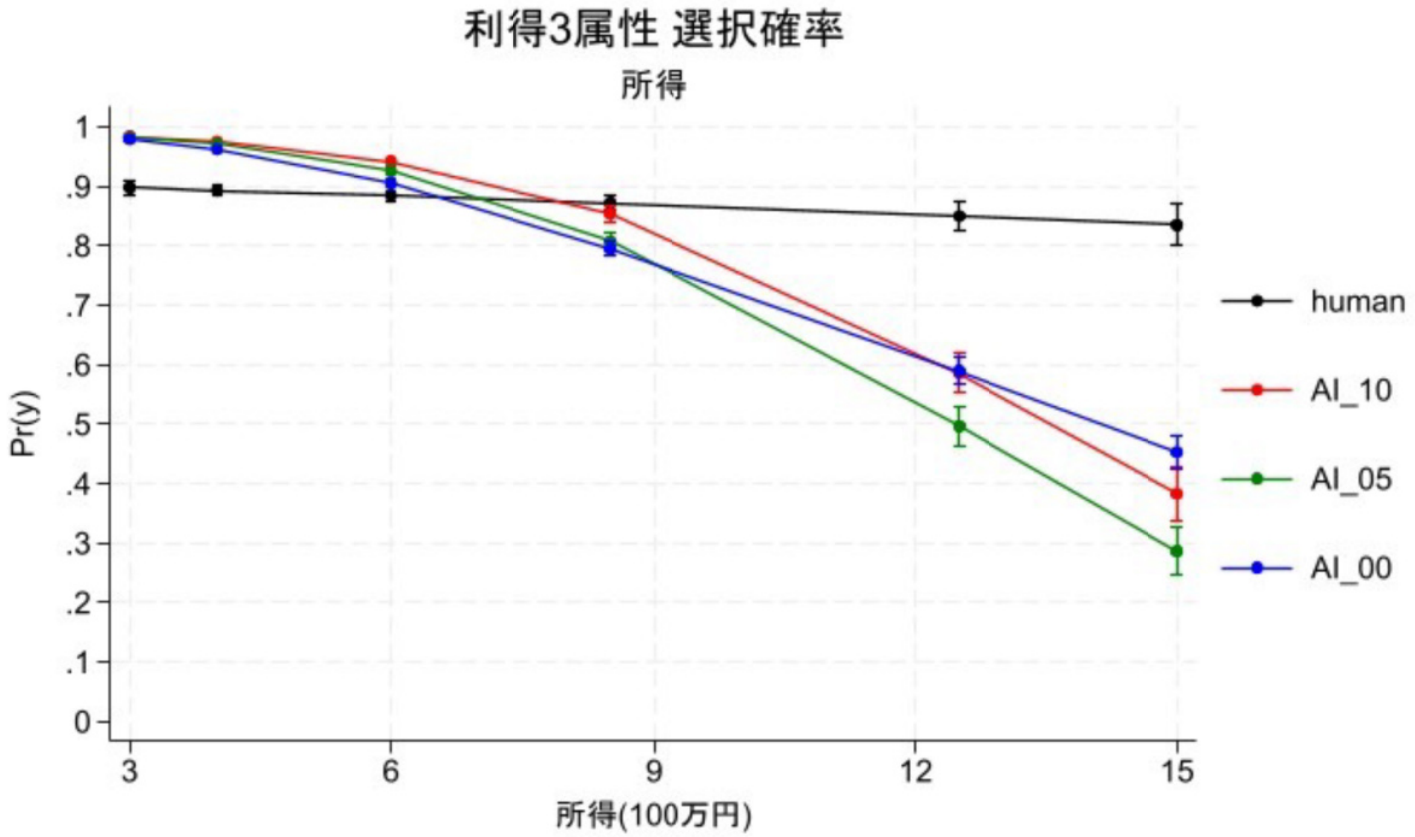
【利得】性別と確実選択肢①の選択確率



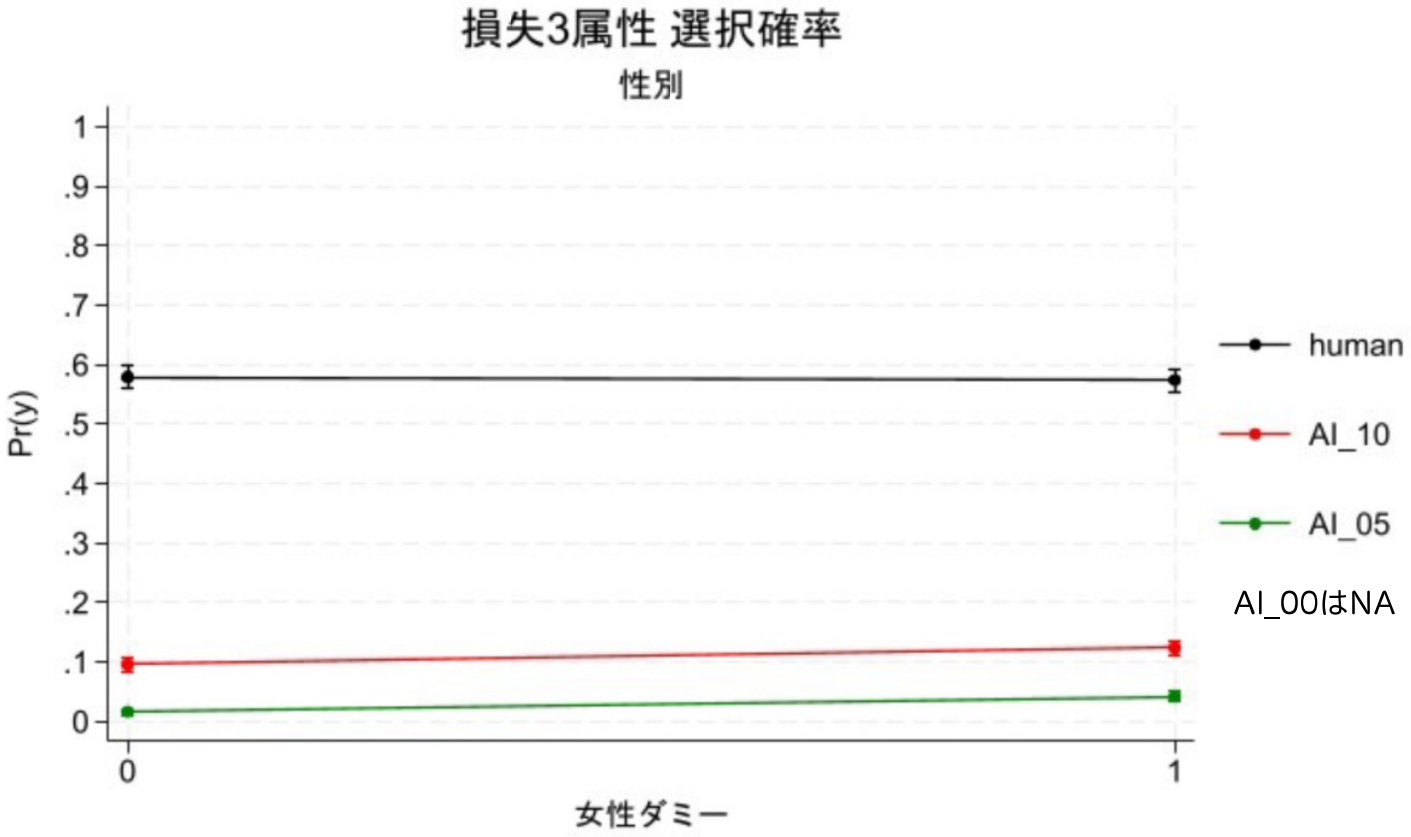
【利得】年齢と確実選択肢①の選択確率



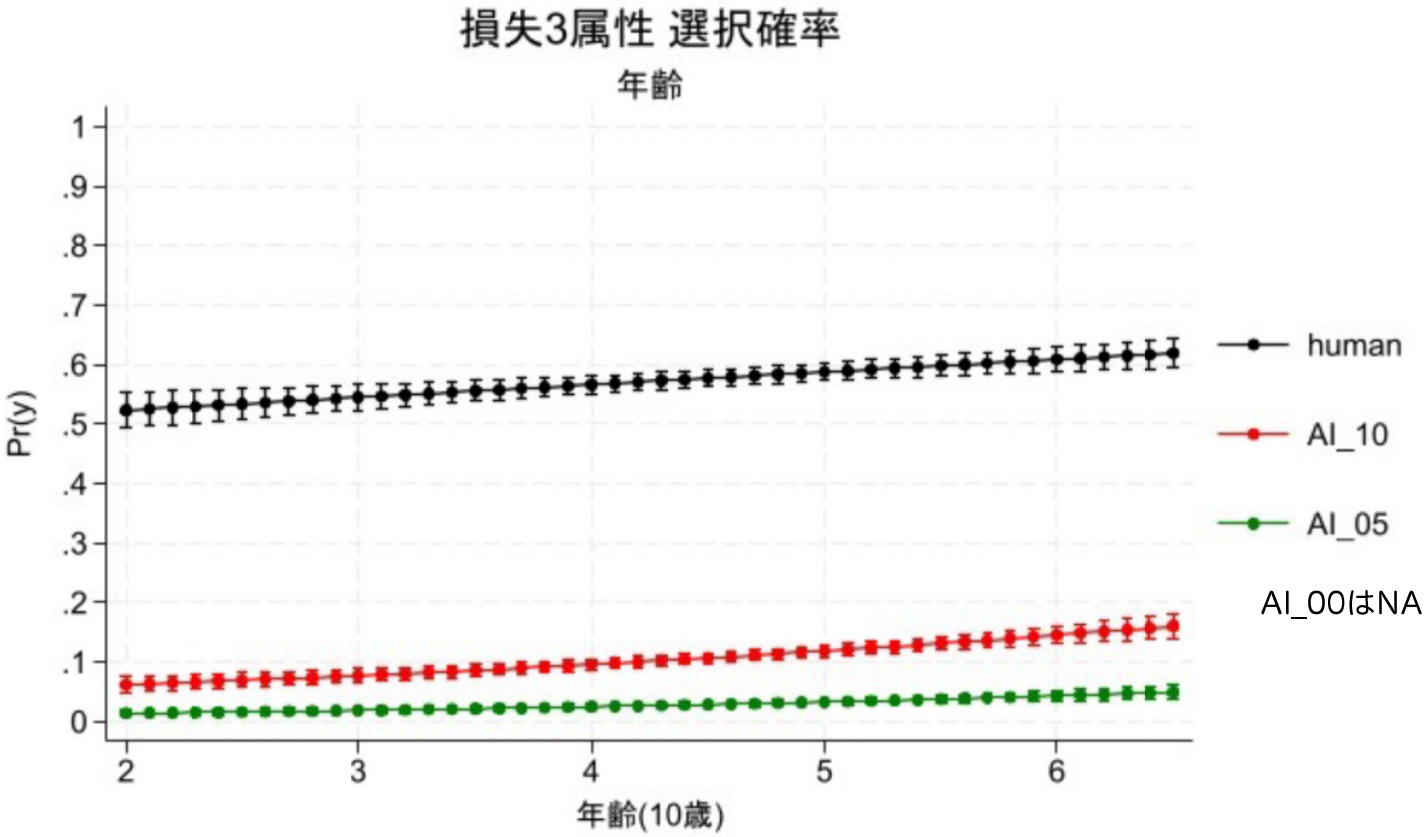
【利得】所得と確実選択肢①の選択確率



【損失】性別と確実選択肢①の選択確率

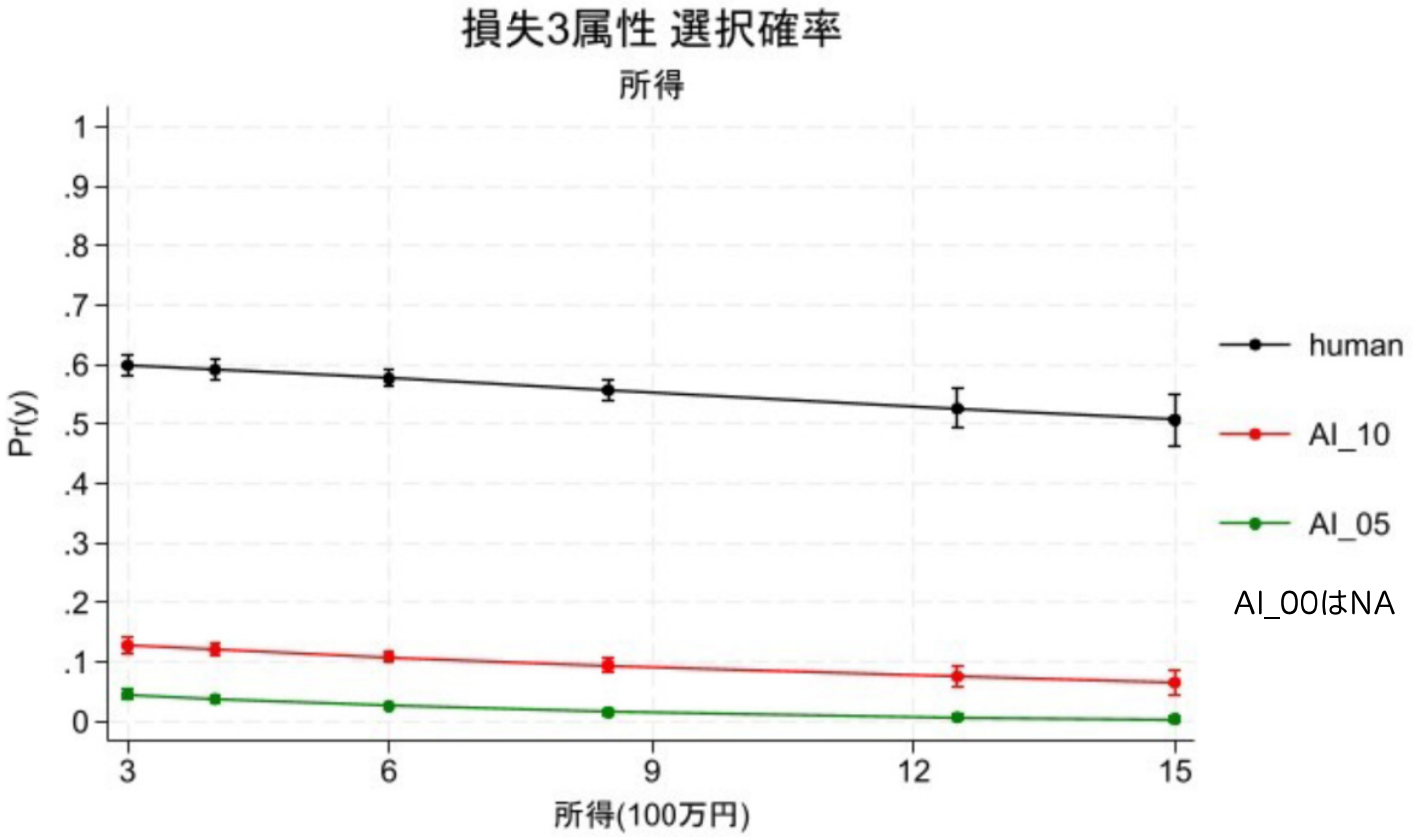


【損失】年齢と確実選択肢①の選択確率



AI_00はNA

【損失】所得と確実選択肢①の選択確率



AI_00はNA

損失回避性のファインチューニング

- 人間の選択確率を訓練データとして「ファインチューニング」。
  同じ「ペルソナ」を持つAI相手に選択確率を得る。
  学習後AIデータをロジット回帰して選択確率を出力。
- 人間の選択確率、学習前後AIの選択確率を比較する。
  選択確率の分布の距離を「ワッサースタイン距離」で測る。
  人間と学習前後AIの損失回避性、限界効果、正解率を比較。
- 学習後AIを人間と学習前AIの選択確率の混合として出力。
  理想的選択を人間と学習前後AIの選択確率の混合として表現。
  →人間型AIパートナー「ヒューマノイド」を開発する。





【利得】確実選択肢を選ぶ比率の比較

- チューニング学習後のAIの選択確率は96%
- 学習前AIと人間と比較して、全体の傾向は維持
- 学習後AIの選択確率は、一部、人間データを過学習

学習前生成AI		
平均確率	0.911	***
	(0.004)	
女性	0.033	***
	(0.007)	
年齢	0.057	***
(10歳)	(0.003)	
世帯所得	-0.027	***
(100万円)	(0.001)	
McFadden R^2	0.347	

学習後生成AI		
平均確率	0.961	***
	(0.003)	
女性	0.019	***
	(0.005)	
年齢	0.030	***
(10歳)	(0.002)	
世帯所得	-0.014	***
(100万円)	(0.001)	
McFadden R^2	0.450	

人間		
平均確率	0.884	***
	(0.005)	
女性	0.041	***
	(0.009)	
年齢	0.015	***
(10歳)	(0.003)	
世帯所得	-0.005	***
(100万円)	(0.001)	
McFadden R^2	0.014	

【利得】属性の回帰係数の比較

- 学習後AIの限界効果は女性：2%、年齢：3%、所得：−1%
- 係数はいずれも予想された符号で統計的に有意
- いずれの効果は学習前AI・人間の間に収まる

学習前生成AI			学習後生成AI			人間		
平均確率	0.911	***	平均確率	0.961	***	平均確率	0.884	***
	(0.004)			(0.003)			(0.005)	
女性	0.033	***	女性	0.019	***	女性	0.041	***
	(0.007)			(0.005)			(0.009)	
年齢	0.057	***	年齢	0.030	***	年齢	0.015	***
(10歳)	(0.003)		(10歳)	(0.002)		(10歳)	(0.003)	
世帯所得	-0.027	***	世帯所得	-0.014	***	世帯所得	-0.005	***
(100万円)	(0.001)		(100万円)	(0.001)		(100万円)	(0.001)	
McFadden R^2	0.347		McFadden R^2	0.450		McFadden R^2	0.0140	

【損失】確実選択肢を選ぶ比率の比較

- チューニング学習後のAIの選択確率は40%
- 学習前AIと人間と比較して、人間の傾向に近づいている
- 学習後AIの選択確率は、学習前AIと人間の間に収まる

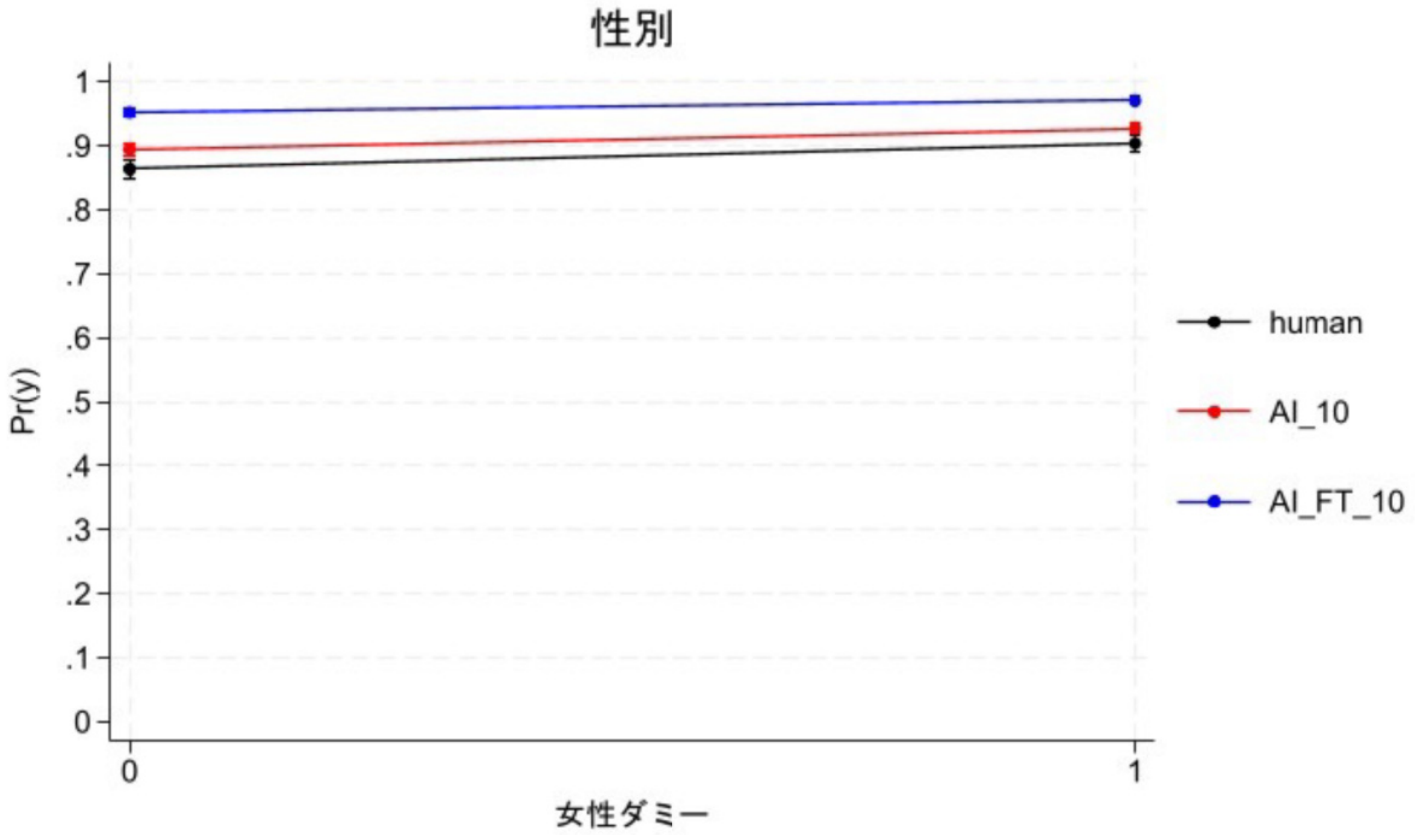
学習前生成AI			学習後生成AI			人間		
平均確率	0.110	***	平均確率	0.404	***	平均確率	0.577	***
	(0.004)			(0.007)			(0.007)	
女性	0.028	***	女性	0.264	***	女性	-0.006	
	(0.009)			(0.011)			(0.014)	
年齢	0.023	***	年齢	0.133	***	年齢	0.021	***
(10歳)	(0.004)		(10歳)	(0.004)		(10歳)	(0.005)	
世帯所得	-0.006	***	世帯所得	-0.003		世帯所得	-0.008	***
(100万円)	(0.002)		(100万円)	(0.002)		(100万円)	(0.002)	
McFadden R^2	0.020		McFadden R^2	0.1804		McFadden R^2	0.004	

【損失】属性の回帰係数の比較

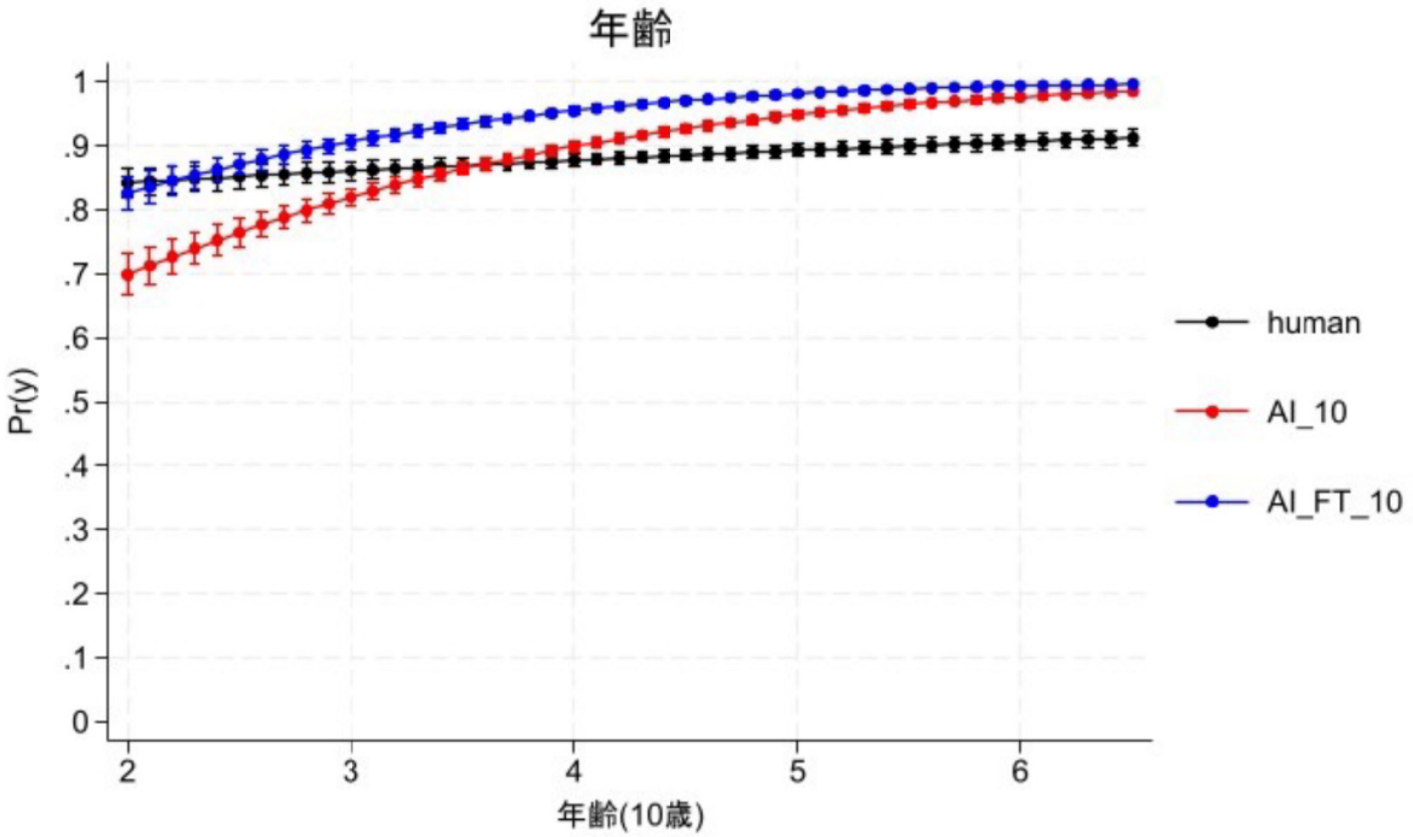
- 学習後AIの限界効果は女性：26%、年齢：13%、所得：0%
- 年齢・女性の係数は予想された符号で統計的に有意
- 年齢・女性の係数は学習前AIと人間を大きく上回る
- 所得の係数は学習前AIと人間をやや下回る

学習前生成AI			学習後生成AI			人間		
平均確率	0.110	***	平均確率	0.404	***	平均確率	0.577	***
	(0.004)			(0.007)			(0.007)	
女性	0.028	***	女性	0.264	***	女性	-0.006	
	(0.009)			(0.011)			(0.014)	
年齢	0.023	***	年齢	0.133	***	年齢	0.021	***
(10歳)	(0.004)		(10歳)	(0.004)		(10歳)	(0.005)	
世帯所得	-0.006	***	世帯所得	-0.003		世帯所得	-0.008	***
(100万円)	(0.002)		(100万円)	(0.002)		(100万円)	(0.002)	
McFadden R^2	0.020		McFadden R^2	0.1804		McFadden R^2	0.0038	

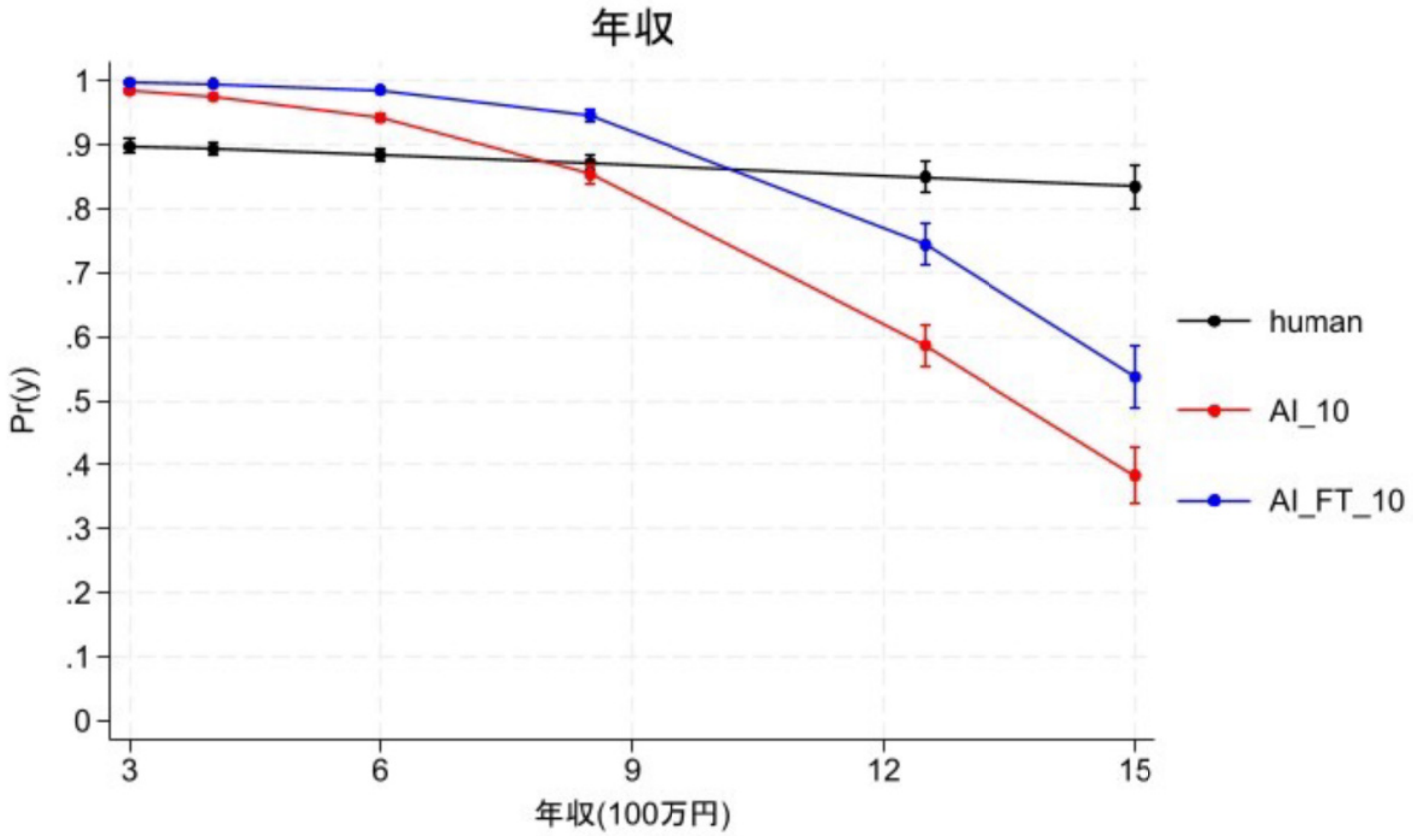
【利得】性別と確実選択肢①の選択確率



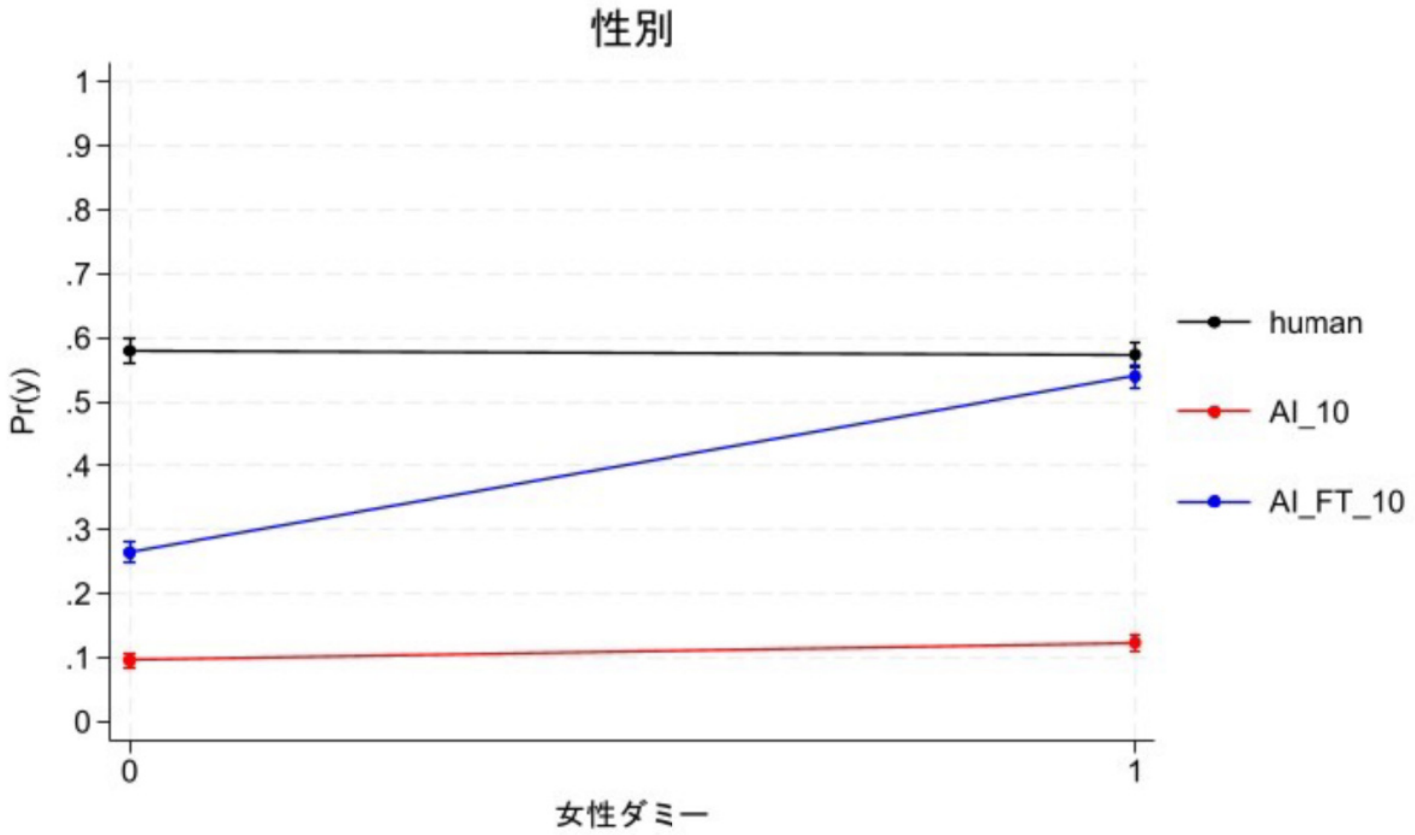
【利得】年齢と確実選択肢①の選択確率



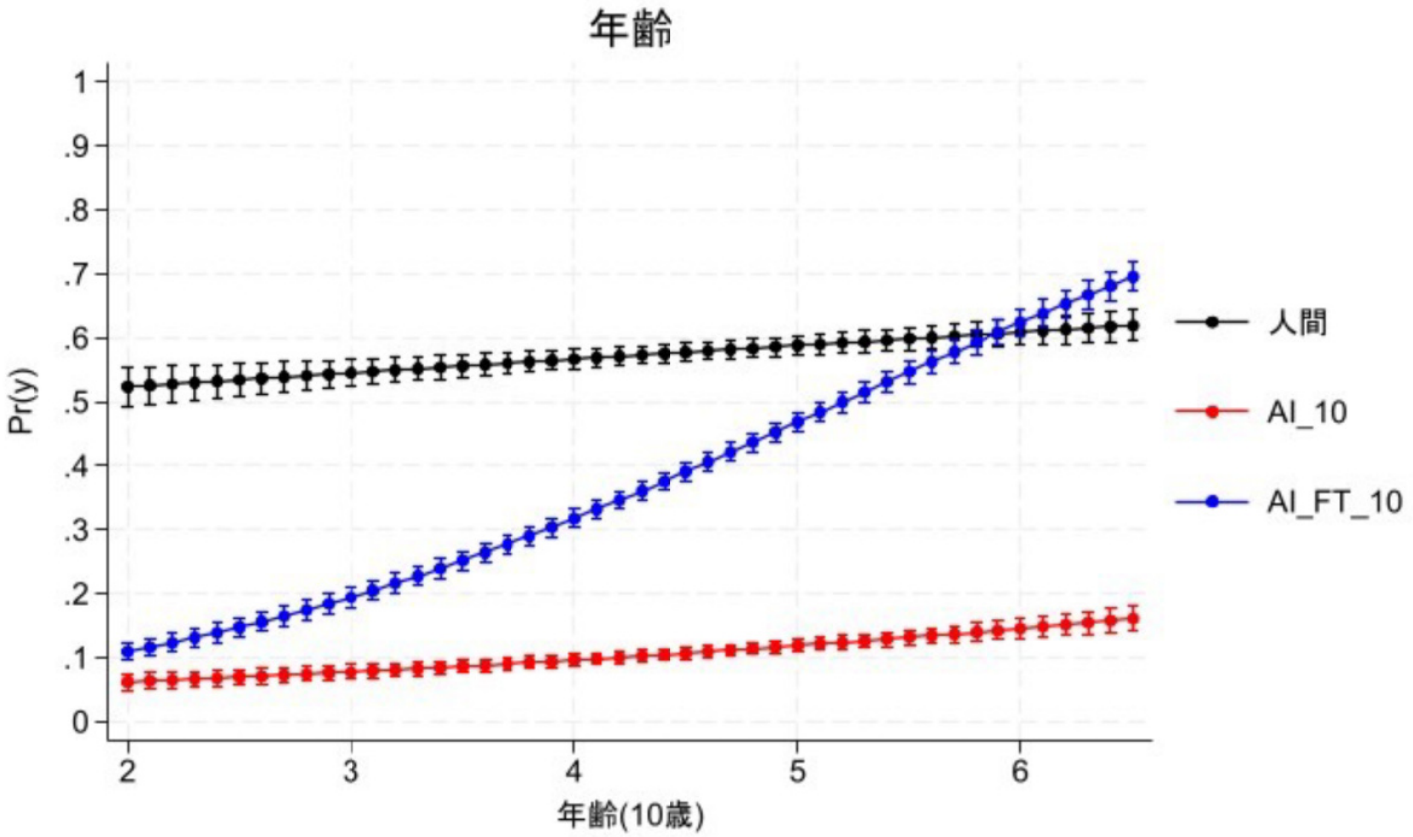
【利得】所得と確実選択肢①の選択確率



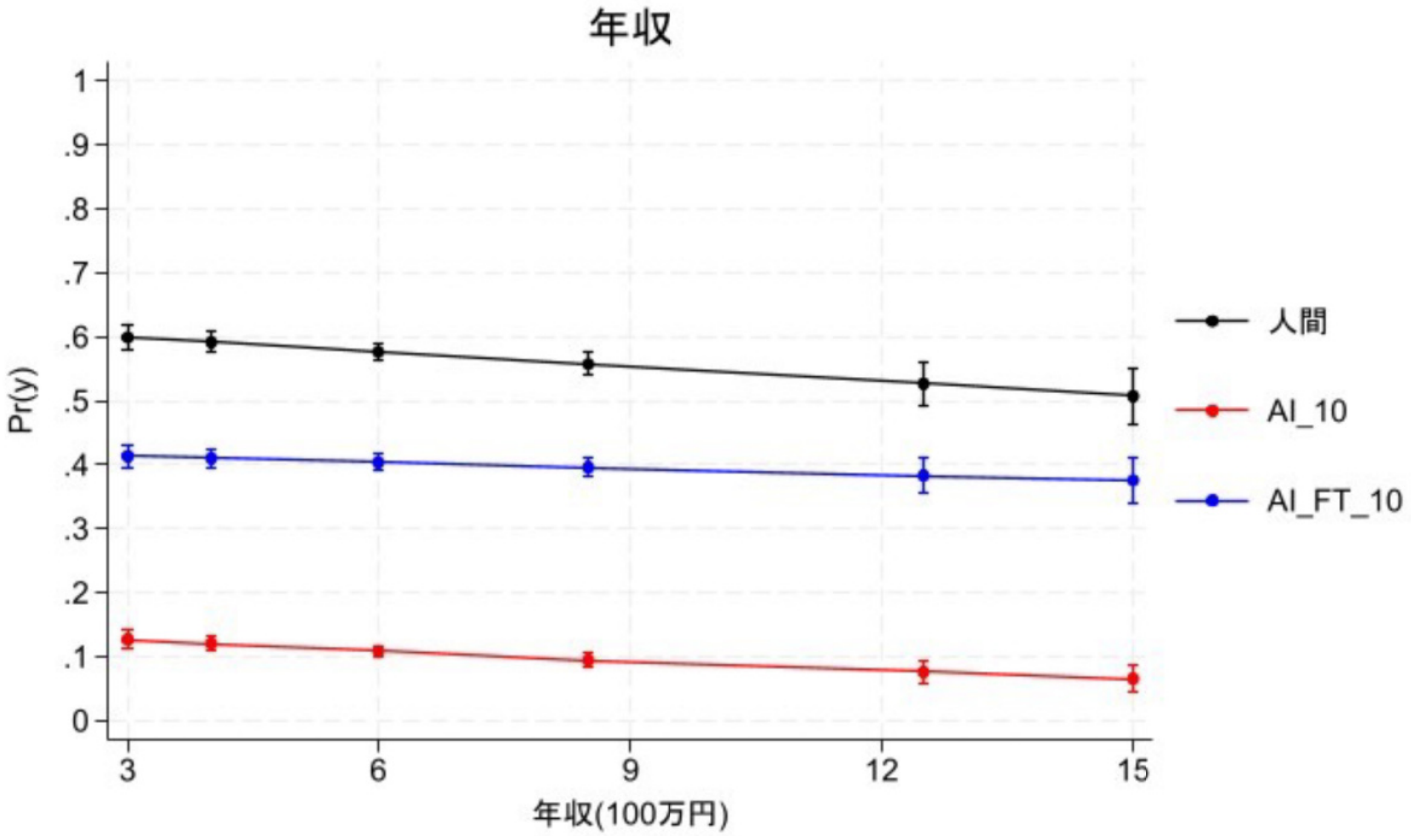
【損失】性別と確実選択肢①の選択確率



【損失】年齢と確実選択肢①の選択確率

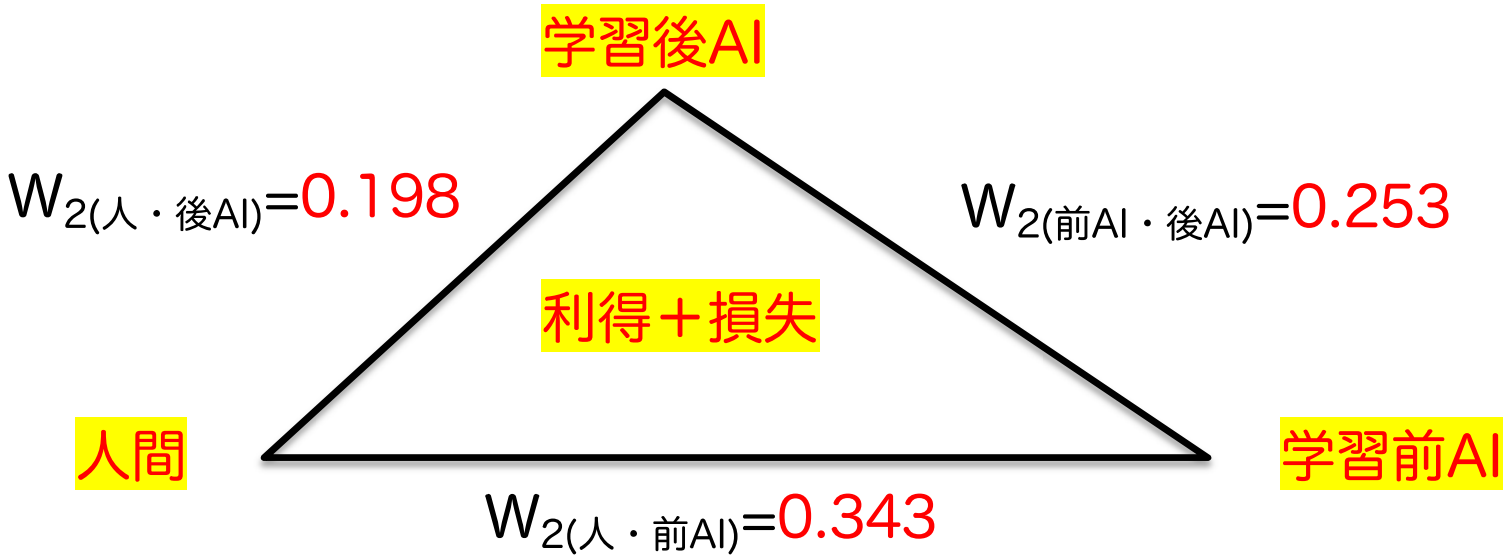


【損失】所得と確実選択肢①の選択確率



【利得＋損失】AIの人間からの損失回避性の学習

- 学習後AIは人間と学習前AIを混合した擬人的「ヒューマノイド」
- 「2次ワッサースタイン距離(W_2)」を用いて分布の距離を測定
- 人間と学習前AIの選択確率分布の距離0.343
- 人間と学習後AIの選択確率分布の距離0.198
- 学習前AIと学習後AIの選択確率分布の距離0.253
- チューニングによりAIと人間の距離は半減した(▲42%)
- 学習後AIは学習前AIと比べて約22%人間寄り



【展望】この研究で行ったことの意義の考察

- 古典的AIは、ルールに従う「記号操作」を使っていたが、「組み合わせ爆発」により複雑な問題に対応できず行き詰まった。
- 現代的AIは、機械学習で高い性能を発揮しているが、意味理解の「記号接地」は解決できていない。
- 出力の偏りを表す「アルゴリズム・バイアス」の課題も残る。
- 本研究では、人間とAIのバイアスを、属性の「ペルソナ」を揃えて定量的に比較する一連の「メトリクス」を考案した。
- AIに人間データを「チューニング」させて、人間とAIの近接距離を測ることで記号接地の問題を迂回した。
- 学習後AIを擬人的に「ヒューマノイド」と呼び、人間とAIの互換性に向けた新しい可能性を示した。
- 本研究で考案した評価軸は、あらゆる人間行動に適用可能であり、「人間AI比較行動科学」という学問分野を創出する。

ご清聴有り難うございました

依田 高典

京都大学大学院経済学研究科 教授



1965 年、新潟県生まれ。
1989 年、京都大学経済学部卒、
1995 年、京都大学大学院経済学研究科博士課程修了。博士(経済学)。現在、京都大学大学院経済学研究科教授。同研究科長・同学部長(2021~2023 年度)。同研究特別教授(2025 年度)。その間、イリノイ大学、ケンブリッジ大学、カリフォルニア大学客員研究員を歴任。専門は応用ミクロ経済学・応用実証経済学。行動経済学・フィールド実験・因果推論ならびに機械学習の融合研究に取り組む。主な著書に『データサイエンスの経済学』(岩波書店)、『スマートグリッド・エコノミクス』(有斐閣)、『ブロードバンド・エコノミクス』(日本経済新聞出版社)、『行動経済学』(中公新書)、『「ココロ」の経済学』(ちくま新書)等がある。日本学術振興会賞、日本行動経済学会ヤフー論文賞、日本応用経済学会学会賞、大川財団出版賞、ドコモモバイルサイエンス奨励賞等を受賞。