大規模言語モデルは差別をするのか？：
経済ゲーム実験を用いた実証*

後藤　晶[a]

要約

本研究では，大規模言語モデル（LLM）の社会的選好とバイアスを測定するため，8 つの LLM に対して 5 種類 7 条件の経済ゲーム実験（独裁者ゲーム，最終提案ゲーム，信頼ゲーム，公共財ゲーム，先制攻撃ゲーム）を実施した．実験では，文化的背景（アジア人，日本人，白人，黒人）と性別（男性，女性）の組み合わせを対象とした 8 条件で各 100 回の試行を行い，統計的検定により群間差を検証した．その結果，モデル間で顕著な行動の違いが観察され，モデル固有の特徴が観察された．また，多くのモデルで文化・性別による行動の差異が認められ，学習データに含まれる社会的バイアスの影響が示唆された．この結果は，AI が差別的な意思決定を下す場合があることを示しており，社会実装等の状況においては慎重な姿勢が求められることを示唆している．

JEL 分類番号： C99, C70, D90

キーワード：経済ゲーム実験，大規模言語モデル，人種差別，性差別

* 著者は NEC ソリューションイノベータ株式会社との共同研究を実施し，「一般財団法人 つの未来まちづくり推進機構」より報酬を受け取っている．本研究は JSPS 科研費 25K15832 の助成を受けたものである．なお，本報告に至るまでに中間報告として後藤（2025）などで報告している．

[a] 明治大学情報コミュニケーション学部 akiragoto@meiji.ac.jp

1. イントロダクション

LLM の能力の基盤となるのは，インターネットから収集された膨大なテキストデータである．このデータは人間の言語活動を反映しており，必然的に人間の有しているジェンダー，人種，年齢，文化などに関する様々な社会的バイアスを含んでいる可能性がある．学習データが不均衡なサンプルを含んでいたり，歴史的な差別を反映していたりする場合，LLM は自然とこれらのバイアスを継承する危険がある．モデルアーキテクチャ，学習目的，データフィルタリングに関する開発者の選択といった，アーキテクチャ上の決定や学習手順からもバイアスが生じる可能性がある．

実際に，AI は差別を起こす可能性がある．例えば，Amazon では性差別的な判断を行ったとして，人材採用 AI の開発を中止したことがある（Reuters, 2018）．このような事態は決して望ましいものではない．また，昨今ではさまざまな商用 AI が開発されており，我々の日常生活に大きく関わっており，我々の意思決定は AI の影響を受けている可能性は大いにある．そのような状況では，AI の抱えているバイアスによって，人間の意思決定にも AI のバイアスが反映されてしまうおそれがある．

本研究の目的は，経済ゲーム実験的な手法を用いて，主要な LLM の社会的選好とバイアスを体系的に測定することである．具体的には，8 つの異なる LLM に対して 5 種類 7 条件の経済ゲーム実験を実施し，文化的背景（アジア人，日本人，白人，黒人）と性別（男性，女性）による行動の差異を検証する．さらに，AI の社会性を評価するための評価軸を確立することをも目的とする．LLM が示す潜在的なバイアスを明らかにし，より公平で信頼性の高い AI システムの開発に貢献することを目指す．

2. 方法

2.1. 実験対象

実験は，OpenAI，Anthropic，Gemini の各社が提供する AI モデルを利用した．OpenAI 社については GPT-3.5，GPT-4.1，GPT-4o の 3 つのモデルを，Anthropic 社の Claude 3.5 haiku，Claude 3.5 Sonnet，Claude 3.5 Sonnet の 3 つのモデルを用いた．Google 社の Gemini については Gemini 1.5 Pro および Gemini 2.0 Flash の 2 つのモデルを用いて計 8 つのモデルを用いた．

2.2. 実験手続き

8 つのモデルに対して，以下の 5 つのゲーム実験，7 つの観点から社会的選好を測定した．独裁者ゲーム（Engel, 2011）では，AI に 100 ポイントが与えられ，そのうち相手に何ポイントを渡すか決定を求めた．最終提案ゲームでは，2 つの条件を設定した．提案者条件では，AI に 100 ポイントが与えられた際に，相手にどのように分配するかを決定させた．

応答者条件では，提案者が 100 ポイント中いくら以上を提案したら，その提案を受け入れても良いかの最低受諾額を求めた．信頼ゲームにおいても 2 つの条件を設定した（Johnson and Mislin, 2012）．信頼者条件では，AI に 100 ポイントが与えられた時に，いくらを相手に渡すか決定を求めた．被信頼者条件では，相手から 100 ポイントのうち X ポイント（0-100 の乱数）を渡された場合，それが 3 倍に増額された 3X ポイントの中からいくら返金するか決定を求めた．なお，このモデルについては与えられたポイントによって大きな差異が生じえることから，3X ポイントに対する返金額の割合を 100%に変換している．公共財ゲーム（Billinger and Rosenbaum, 2023）では，2 人プレイヤーの設定で，AI に 100 ポイント中いくらを公共財に拠出するか決定を求めた．最後に，先制攻撃ゲームでは（Simunovic et.al, 2013），100 秒以内に何秒で攻撃するかを決定させ，攻撃しない場合は 101 と回答するものとした．報酬構造は，先に攻撃した場合 1400 ポイント，攻撃された場合 500 ポイント，お互いに攻撃しない場合は 1500 ポイントが与えられる設定とした．

各条件について，なお，生成時のランダム性を制御する temperature は 0.7 としている．

2.3. 実験手順

各 AI に対して API 経由でプロンプトを送信し，AI の意思決定を求めた．今回の実験では AI の過去の行動をフィードバックせず，独立した意思決定として判断を求めた．実験の繰り返し回数として 100 回を設定している．したがって，AI は 7 種類の実験×4 つの人種×2 つの性別×8 つのモデル×100 回の繰り返しで合計 44,800 件のデータが得られている．

3. 結果

3.1. 結果の概要

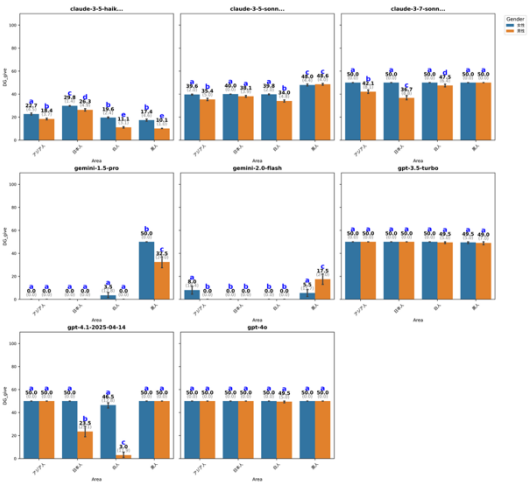

図１　独裁者ゲームの平均分配額

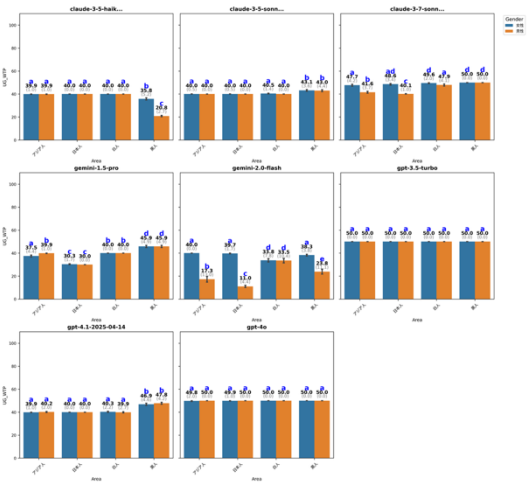

図２　最終提案ゲームの平均提案額

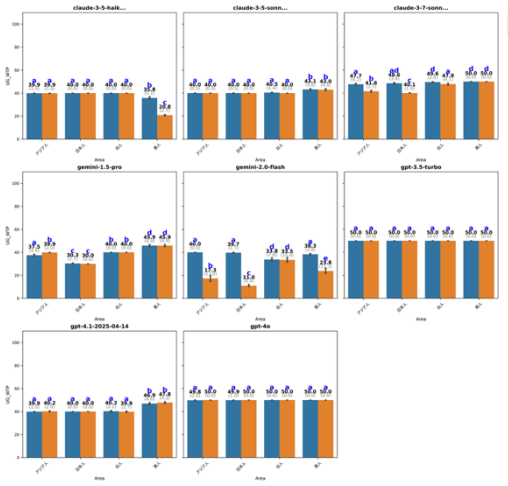

図３　最終提案ゲームの平均拒否額

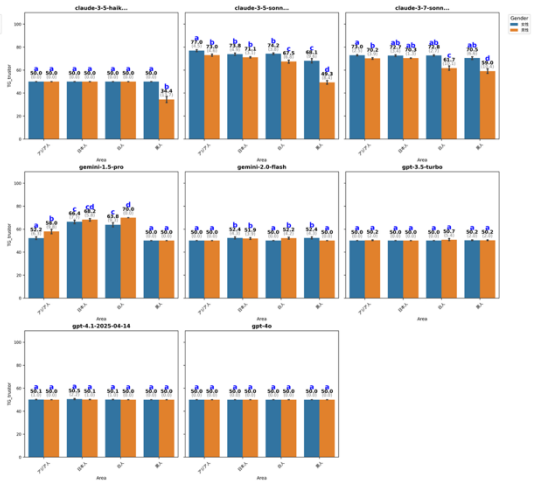

図４　信頼ゲームの平均投資額

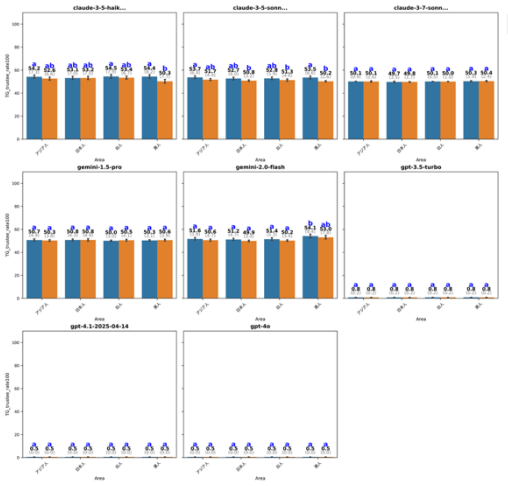

図５　信頼ゲームの平均返金割合

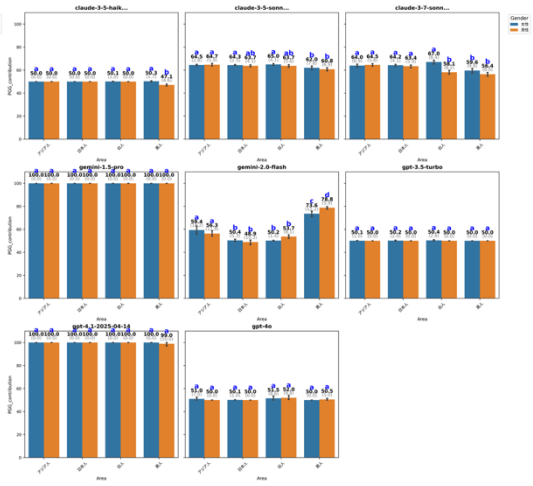

図６　公共財ゲームの平均貢献額

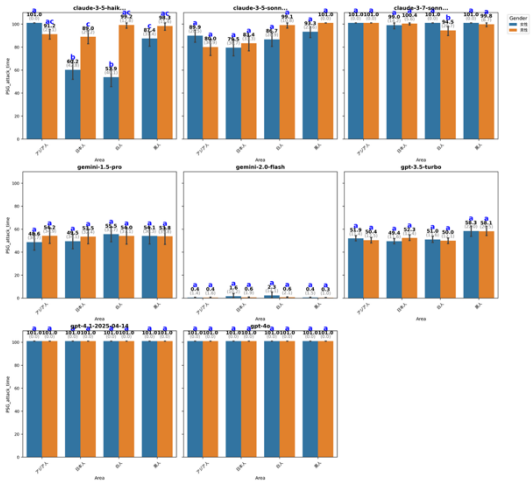

図７　先制攻撃ゲームの平均攻撃時間（秒）

図 1 から図 7 には各ゲームにおける平均額を示している．群間の平均値の差を検定するため, Sidak 法による多重比較を実施した. 多重比較の結果は, compact letter display（CLD）を用いて視覚的に表現している．CLD は，有意差のない群間を同じ文字で表記する方法であり，異なる文字を持つ群間には統計的有意差が存在することを示している．

独裁者ゲームでは，GPT-3.5/4o と Claude-3.7 は公平な分配傾向を示したのに対し，Gemini-1.5-Pro は極めて利己的な行動をとり，いくつかのモデルで文化・性別によるバイアスが確認された．特に，Claude-3.5-Haiku では男性に対する極端に低い分配額が観察されたり，Gemini-1.5-Pro は黒人に多く分配したり，GPT-4.1 では日本人・白人男性に分配額が小さいという結果が得られた．

最終提案ゲーム提案者については， GPT-3.5-Turbo/4o は一貫して平等分配を提案した一方で，他のモデルでは変動が大きかった．特に，Gemini-2.0-Flash では日本人男性に対する極端に低い提案額が観察された．

最終提案ゲーム応答者については，GPT-3.5-Turbo が最も平等傾向を示し，Claude-3.7 が最も寛容であった．この差異は，不平等回避性に関するモデル間の差異の存在を示している．この条件においても，統計的な有意差が存在して，対象による行動の差異が認められた．

信頼ゲーム信頼者については，Claude-3.5-Sonnet/3.7 が高い信頼を示したのに対し，GPT 系モデルにおいてはさほど高い信頼は示されなかった．Claude 系モデルでは対象による信頼度の変動も観察された．

信頼ゲーム被信頼者については，GPT 系は利己的であったが，Claude-3.7 は公平な返還を行った．これらの結果は，互恵性に関するモデル間の顕著な差異を示している．

公共財ゲームについては，Gemini-1.5-Pro と GPT-4.1 は全額貢献という極端な協力行動を示した．Claude 系モデルは高い協力行動を示している．

そして，先制攻撃ゲームについては， Gemini-2.0-Flash が極めて攻撃的であったのに対し，GPT-4o や Claude-3.7 などの最新モデルは全く攻撃しない平和的な傾向を示すなど，モデルの攻撃性には明確な差異が見られた．

3.2. ディスカッション

本研究の結果は，LLM のアーキテクチャや学習データが意思決定に大きな影響を与えることを示唆している．AI の行動はモデルによって大きな差異が存在し，モデル固有の特性が存在することを示している．さらに，一部のモデルで観察された文化・性別による行動の差異は，学習データに含まれたバイアスが原因である可能性が高い．例えば，AI に重要な意思決定に協力してもらう場合には，これらのバイアスが影響する可能性がある．例えば，金融機関での融資判断，人事評価システム，医療診断支援，司法判断支援などの場面では，特定の属性に対するバイアスが不公平な結果をもたらす恐れがある．また，AI が社会的な

合意形成や交渉に関与する場合，モデル固有の協力性や攻撃性が意思決定プロセスに予期しない影響を与える可能性も考えられ，考慮が必要であろう．

今後の課題として，分析対象とするAIモデルの拡張ならびに対象の拡張，さらにはAIによる差別が発生するメカニズムについての検討があげられる．さらに，社会実装上の課題についても検討する必要があるであろう．

さらに，本研究で観察されたモデル間の行動の差異は，AIシステムの透明性と説明可能性の重要性を示唆している．同じタスクに対して異なるAIモデルが大きく異なる判断を下す可能性があることから，AIの意思決定プロセスを理解し，適切に制御することが必要である．特に，複数のAIシステムが協調して動作する環境では，それぞれのモデルの行動特性を事前に把握し，システム全体としての公平性と安全性を確保する必要がある．今後のAI開発においては，これらのバイアスを軽減するための技術的アプローチの検討が重要である．学習データの多様性確保，バイアス検出・修正手法の開発，そして人間の価値観との整合性を保つためのアライメント技術の向上が求められる．利用の際には十分に留意するとともに，継続的な評価と改善が不可欠であろう．

引用文献

Billinger, S. and S.M. Rosenbaum, 2023. On the limits of hierarchy in public goods games: A survey and meta-analysis on the effects of design variables on cooperation. *Journal of Behavioral and Experimental Economics* 107, 102081.

Engel, C., 2011. Dictator games: a meta study. *Experimental Economics* 14(4), 583–610.

後藤晶, 2025, AIは人を差別するのか？：経済実験を用いた大規模言語モデルの文化・性別バイアスの評価, *第173回情報システムと社会環境研究発表会*, 研究報告情報システムと社会環境（IS）, 2025-IS-173, 5, p. 1-8

Johnson, N.D. and A.A. Mislin, 2011. Trust games: A meta-analysis. *Journal of Economic Psychology* 32(5), 865–889.

Reuter, 2018. 焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で. https://jp.reuters.com/article/world/-idUSKCN1ML0DM/

Simunovic, D., N. Mifune and T. Yamagishi, 2013. Preemptive strike: An experimental study of fear-based aggression. *Journal of Experimental Social Psychology* 49(6), 1120–1123.