大規模言語モデルにおける人種・ジェンダーバイアスの定量的評価：

経済ゲーム実験を用いた実証分析

Quantitative Assessment of Race and Gender Bias in Large Language Models:

An Empirical Analysis Using Economic Game Experiments

後藤　晶*[1]
Akira GOTO
*[1] 明治大学　Meiji University

要旨: 本研究は，AWS Bedrock 経由で取得した 13 種類の大規模言語モデル（LLM）における人種・ジェンダーバイアスを，行動経済学ゲームを用いて定量的に評価した．独裁者ゲーム，最後通牒ゲーム，信頼ゲーム，公共財ゲームの 4 種類のゲームにおいて，6 つの人種（日本人，アジア人，白人，黒人，ヒスパニック，中東系）と 3 つのジェンダー（男性，女性，ノンバイナリー）の組み合わせによる意思決定を分析した．

二要因分散分析（ANOVA），線形回帰分析，推定周辺平均（EMM）による多重比較を実施した結果，モデル間で人種・ジェンダー要因に対する反応パターンに有意な差異が認められた．特に，特定のモデルにおいて人種間の利他性や信頼性に統計的に有意な差が観察され，LLM が学習データに内在するバイアスを反映している可能性が示唆された．本研究は，AI 倫理と公平性の観点から，LLM の社会実装における課題を提示する．

キーワード: 経済ゲーム実験，LLM，社会性

Abstract: This study quantitatively assessed race and gender bias in 13 large language models (LLMs) accessed via AWS Bedrock using behavioral economics games. We analyzed decision-making across combinations of six races (Japanese, Asian, White, Black, Hispanic, Middle Eastern) and three genders (male, female, non-binary) in four games: Dictator Game, Ultimatum Game, Trust Game, and Public Goods Game.

Two-way ANOVA, linear regression, and multiple comparisons using estimated marginal means (EMM) revealed significant differences in response patterns across models by race and gender. Notably, statistically significant differences in altruism and trustworthiness among races were observed in specific models, suggesting that LLMs may reflect biases inherent in their training data. This study highlights challenges in the social implementation of LLMs from the perspectives of AI ethics and fairness.

Keywords: Economic Game Experiment, LLM, Sociality

1. 問題

大規模言語モデル（LLM）の社会実装が急速に進展する中，これらのモデルが示すバイアスの評価と軽減が重要課題となっている．既存のバイアス研究では，感情分析など，テキストマイニングおける偏りが主に分析されてきたが（Wang et.al, 2018），協力・公平・信頼といった社会的行動の文脈におけるバイアスは十分に検討されていない．LLM が採用支援や融資審査など人間の社会的行動に直接影響を与える場面で使用されるようになるにつれ，属性情報に基づく行動の偏りは実質的な影響を及ぼす可能性がある．

実験経済学では，独裁者ゲームや信頼ゲームといった標準的な経済ゲームを用いて，人間の社会的選好を定量的に測定してきた．これらのゲームは，寛大さ，公平性，信頼，協力といった社会的

行動の基本的側面を測定する行動経済学の標準的ツールである．人間被験者を対象とした既存研究では，経済ゲームにおける行動が人種やジェンダーにより異なることが報告されている．

近年，LLM が人間に類似した推論能力を示すことから，経済ゲーム実験を LLM に適用する試みが始まっている．しかし，既存研究は限定的なモデルを対象としており，多様なモデルにおける属性依存性を包括的に比較した研究はほとんど行われていない（Mei et al. 2024, Horton, 2023, Xie et al., 2024）．特に，複数のモデル・複数のゲームタイプを横断的に比較し，属性効果の一般化可能性を検証した研究は存在しない．

本研究では，AWS Bedrock 提供の 13 種類の LLM に対して，人種（6 カテゴリ）とジェンダー（3 カテゴリ）を操作した 7 種類の経済ゲーム実験を実施し，総計 163,800 サンプルを収集した．本研究の貢献は，(1)多様なモデルの包括的比較，(2)複数の社会的行動次元の測定，(3)大規模サンプルによる統計的に頑健な評価，である．

2. 方法

2.1. 実験対象と条件

AWS Bedrock API を用いて 13 種類の LLM を実験対象とした（表 1）．人種は 6 カテゴリ（白人，黒人，アジア人，日本人，ヒスパニック，中東系），ジェンダーは 3 カテゴリ（男性，女性，ノンバイナリー）を設定し，合計 18 条件を設けた．プロンプトには「相手は[人種][ジェンダー]です」という形式で属性情報を提示した．

今回，用いたモデルは Cohere 系の Command-R，Command-R-Plus，Amazon 系の Nova-Lite，Nova-Micro，Nova- Pro，Titan-Text-Express，Meta が開発した Llama-3-70B，Mistral 系の Mistral-7B，Mistral-Large，Mistral-Small，Mixtral-8x7B，Alibaba 系の Qwen3-Coder-30B である．

2.2. 経済ゲームの概要

標準的な経済ゲーム実験のプロトコルに従い，後藤（in print）の枠組みを用いて 7 種類の経済ゲームを実施した．全ゲームで初期保有額を 100 ポイントとし，各条件で 100 回の繰り返し試行を実施した（$n = 100$）．**(1) 独裁者ゲーム**：参加者が 100 ポイントを保有し，匿名の相手への配分額を一方的に決定する．このゲームは純粋な利他性や寛大さを測定する．**(2) 最終提案ゲーム・提案者**：参加者が 100 ポイントの分配案を提示し，相手が受諾・拒否を決定する．拒否された場合は両者とも 0 ポイントとなる．提案者の行動は公平性への配慮と戦略的推論を反映する．**(3) 最終提案ゲーム・応答者**：提案者から 100 ポイントの中からいくら受け取ったら拒否しないかという WTA を尋ねた．このゲームは不平等回避や負の互恵性を測定する．**(4) 信頼ゲーム・投資者**：参加者が 100 ポイントから投資額を決定し，投資額は 3 倍に増幅されて受託者に渡される．このゲームは信頼を測定する．**(5) 信頼ゲーム・受託者**：投資者から 0-100 ポイントの中のランダムなポイントを受け取った受託者が返礼額を決定する．ここでは特に受託者が受け取ったポイントに対する返礼額の割合を分析対象としている．このゲームは信頼性や正の互恵性を測定する．**(6) 公共財ゲーム**：2 人グループで各自が 100 ポイントから公共財への貢献額を決定する．貢献総額の 1.6 倍が全員に等分配される．このゲームは協力を測定する．**(7) 先制攻撃ゲーム**：2 人が同時に 100 秒中何秒で攻撃をするか，もしくは攻撃をしないかを選択する．両者が攻撃をしなければ各 1500 ポイント，一方，攻撃をする場合は先に攻撃をしたプレイヤーが 1400 ポイント，攻撃された側が 500 ポイントとなる．このゲームは紛争解決や攻撃性を測定する．

各ゲームにおいて，LLM には相手プレイヤーの属性情報（人種・ジェンダー）が提示され，総計 163,800 サンプル（13 モデル×18 条件×7 ゲーム×100 試行）を収集した．temperature 値は 0.7 に固定した．

3. 結果

図 1 に全 13 モデル×7 ゲームの結果を示す．モデル間・ゲーム間で顕著な差異が観察される．Command-R は独裁者ゲームで条件依存性が高く（0.5〜43.4 ポイント），日本人女性条件と日本人男性条件で 87 倍の差異を示した．この結果は，同一モデルでもプロンプトの属性情報により行動が大きく変化することを示している．Nova 系と Command-R-Plus は全ゲーム・全条件で安定しており，属性情報による影響が低い．女性条件（赤）が男性条件（青）より高い傾向が多くのゲームで観察され，学習データに含まれるジェンダーステレオタイプの影響が示唆される．

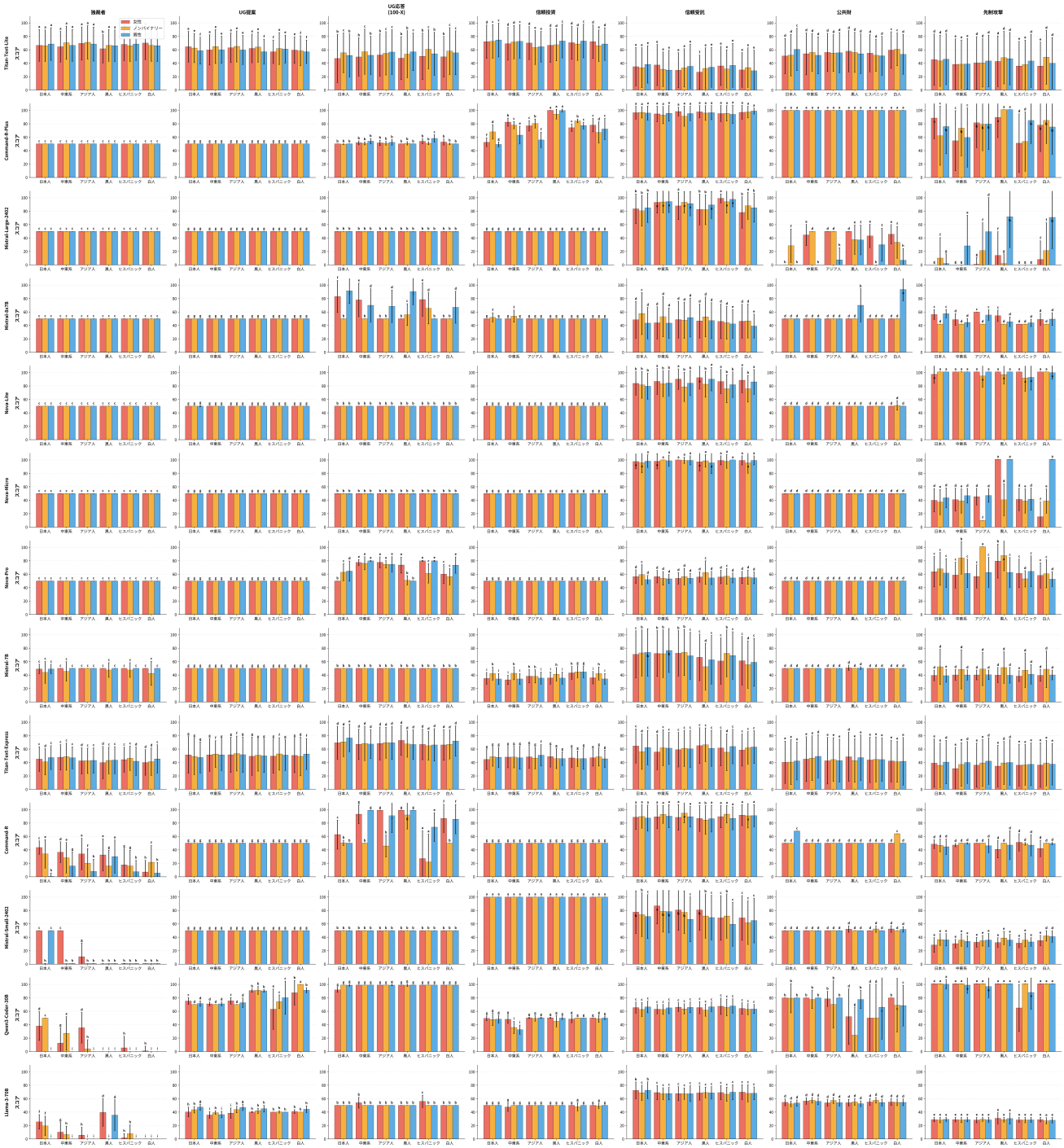

図 1 ***全モデルの人種・性別効果（13 モデル×7 ゲーム，CLD 付き，総計$N = 163,800$）．各行はモデル，各列はゲームを示す．横軸は人種，各人種内で性別（赤=女性，オレンジ=ノンバイナリー，青=男性）をグループ化している．エラーバーは標準偏差（各条件$n = 100$），文字は多重比較の結果を示している．***

続いて，表 2 に二要因 ANOVA の結果を示す．ここでは，5%水準で有意差が認められた条件を示す．全 91 モデル・ゲームの組み合わせ（13 モデル×7 ゲーム）のうち，人種効果は 49 件（54%），ジェンダー効果は 45 件（49%），交互作用は 43 件（47%）で有意であった．特に先制攻撃ゲームでは，13 モデル中 10–11 モデルで有意効果が検出され，属性情報への感応性が最も高かった．

4. 考察

4.1. 主要な知見

本研究は，163,800 サンプルにより，LLM が人種・ジェンダー情報に基づいて顕著に異なる社会的行動を

示すことを実証した．特に，Command-R において，条件間で 87 倍の差が観察されたことは，プロンプトの属性情報がモデル出力に極めて大きな影響を及ぼすことを示す．この結果は，LLM の社会的行動における属性バイアスが，意思決定レベルで実質的な影響を持つことを示している．

検出された人種効果およびジェンダー効果は，学習データのステレオタイプがモデルの振る舞いに反映されている可能性が示唆される．一定程度，人間被験者の実験と一致し，LLM が学習データから社会的ステレオタイプを獲得している可能性が示されている．ただし，全体的な結果を確認すると，必ずしも一貫しているわけではなく，モデルによる差異を含めて，単純な議論ができるわけではなく，複雑な様相を呈していると考えられる．

4.2. インプリケーション

モデル間の多様性は，開発元の設計方針やアライメント手法の違いが社会的選好に反映されている可能性を示している．Nova 系モデルは全条件で公平分配を選択し，明示的な公平性アライメントの存在を示唆する．一方，Command-R や Llama-3-70B は条件により大きく変動し，属性情報への感受性が高く，特に変動が大きなモデルは実用上のリスクがあると言えるであろう．

採用支援や融資審査といった社会的影響の大きい応用においては，モデル選択が重要である．先制攻撃ゲームにおいて最も多くのモデルで有意効果が検出されたことは，紛争解決や交渉支援といった応用において LLM を使用する際の重要な留意点である．

4.3. 限界と今後の課題

本研究の限界として，(1)AWS Bedrock のモデルのみ対象，(2)プロンプト形式固定，(3)人間被験者との直接比較なし，(4)英語プロンプトのみ使用，が挙げられる．今後の課題は，(1)他の主要モデルへの拡張，(2)プロンプト形式や言語の影響の検証，(3)人間被験者との直接比較実験，(4)バイアス軽減手法の開発と評価，(5)実世界のアプリケーションにおける影響評価，である．

5. 結論

本研究は，経済ゲーム実験という定量的手法により，LLM における人種・ジェンダーバイアスを包括的に評価した．AWS Bedrock 提供の 13 種類の LLM に対して 7 種類の経済ゲームを実施し，総計 163,800 サンプルを収集した結果，以下の知見が得られた．

第一に，LLM の社会的行動が属性情報に強く依存することが示された．第二に，回帰分析により人種効果およびジェンダー効果が観察され，学習データのステレオタイプの影響が確認された．第三に，モデル間で大きな多様性が観察され，開発元のアライメント手法の違いが社会的行動に反映されることが示された．

これらの結果は，LLM の社会実装において属性情報の扱いが重要課題であることを示している．特に，採用支援，融資審査，交渉支援といった社会的影響の大きい応用においては，モデル選択やバイアス検出手法の確立が不可欠である．本研究で開発した経済ゲーム実験による評価手法は，LLM の公平性評価の標準的ツールとして活用できる可能性がある．今後，より多様なモデルとプロンプト形式を対象とした研究，およびバイアス軽減手法の開発が求められる．

謝辞

本研究にあたり，科研費補助金 25K15832，ならびに公益財団法人鹿島学術振興財団，公益財団法人電気通信普及財団の支援を受けました．ここに記して感謝申し上げます．

文　　献

後藤　晶.（in print）.“経済ゲーム実験を用いた AI の社会性評価：新たな定量評価指標の提案と検証”, 情報処理学会誌

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.

Mei, Qiaozhu, et al. "A Turing test of whether AI chatbots are behaviorally similar to humans." Proceedings of the National Academy of Sciences 121.9 (2024): e2313925121.

Wang, Alex, et al. "GLUE: A multi-task benchmark and analysis platform for natural language understanding." Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP. 2018.

Xie, Chengxing, et al. "Can large language model agents simulate human trust behavior?." Advances in neural information processing systems 37 (2024): 15674-15729.