最後通牒ゲームの大規模言語モデルを用いたシミュレーション：バイアス補正による実験結果の再現性検証*

北代絢大[a]　深澤祐援[b]　西野成昭[c]

要約

本研究は, 大規模言語モデル（LLM）を用いて自律的に意思決定を行うエージェントを用いたシミュレーションによる, 最後通牒ゲームの経済実験の結果再現に取り組んだ. 近年 LLM から実際の人間のような反応を引き出そうとする研究が社会科学の様々な分野で行われているが, その報告は一貫的でない. 本研究ではその要因として, 実際の人間のものとは異なる LLM の心理的・認知的側面に焦点を当てた. 行動経済学における様々な指標を利用して LLM のバイアスを補正することで, 最後通牒ゲームの実験結果を再現可能かどうか検証した. シミュレーション結果は実際の実験データと完全には一致しないものの, 最後通牒ゲームの実際の実験結果としてよく知られる特徴は再現出来た. 将来的に LLM に実際の人間の反応を模倣させられるようになれば, 実社会における人間を含む系に対するシステムの事前検証が, 実際の人間に見立てた LLM を用いた社会シミュレーションにより可能になるだろう.

JEL 分類番号： C60, C63, C80

キーワード：経済実験, 大規模言語モデル, マルチエージェントシミュレーション, 最後通牒ゲーム

*なお，本論文に関して，開示すべき利益相反関連事項はない．本研究は JSPS 科研費 (22H01710)の助成を受けたものである.

[a] 東京大学大学院工学系研究科　a.kitadai@css.t.u-tokyo.ac.jp

[b] 東京大学大学院工学系研究科　fukasawa@css.t.u-tokyo.ac.jp

[c] 東京大学大学院工学系研究科　nishino@tmi.t.u-tokyo.ac.jp

1. はじめに

大規模言語モデル（LLM）を用いて自律的に意思決定を行うエージェント（生成エージェント）から実際の人間のような反応を引き出そうとする研究が社会科学の様々な分野から注目を集めつつある．実際，生成エージェントと実際の人間の反応を比較する研究が経済学や心理学における被験者実験，市場調査や世論調査においてそれぞれ様々なテーマで行われている（Aher et al., 2023; Li et al., 2024; Argyle et al., 2023; Lee et al., 2024）．しかし，いずれの分野においても両者が一致したという報告と乖離したという報告が混在しており，生成エージェントの実際の人間との代替可能性は未だ明らかになっていない．

先行研究の報告が一貫的でない現況の一因として，LLM 自体が人間とは異なる様々なバイアスを有すという点が考えられる．先行研究では，LLM に対して心理的・認知的テストを実施した際に人間とは異なる結果が得られることが報告されている（Hagendorff, 2023; Hagendorff et al., 2023）．よって，LLM と人間との間に乖離のある心理的・認知的側面が重要となる局面において，LLM と人間の反応にも乖離が生じるのは自然である．一般に，実際の人間の反応は心理的・認知的側面にも影響を受けることが多いため，生成エージェントが人間の反応を模倣するには，LLM に内包される固有のバイアスを補正する必要があるだろう．しかし，LLM に内包される心理的・認知的バイアスを補正することにより，生成エージェントの出力を実際の人間の反応に近づけようという試みは未だ行われていない．

特に経済実験に焦点を当てた最近の研究では，生成エージェントの推論能力が高まるにつれて，シミュレーション結果がナッシュ均衡に近づくことが示されている（Kitadai et al., 2024）．これは，LLM の推論能力が向上していく一方で，実際の被験者を通じた実験結果を生成エージェントで再現するには，更なる工夫が必要であることを示唆している．心理的・認知的バイアスの補正を通じた生成エージェントの特徴づけは，この課題に対する有望なアプローチとなる可能性がある．

本研究の目的は，行動経済学における様々な指標で特徴づけられた生成エージェントを被験者に見立てたシミュレーションによる，実際の経済実験の再現可能性を明らかにすることである．特に，生成エージェントの特徴づけには，行動経済学でよく知られる多様な指標を可能な限り網羅的に測定し，それらの関係を分析した Chapman et al. (2023)の知見及び公開データ（Chapman et al., 2022）を利用する．また実験の題材としては，古典的な経済実験の一つであり人間の非合理的な振る舞いが観察される同時手番の最後通牒ゲームに焦点を当て，Lin et al. (2020) で報告されている人間を被験者とした実験結果と比較する．

本研究は生成エージェントが実際の人間の代替として機能する可能性の解明に大きく寄与する．将来的に，生成エージェントから実際の人間と同様の反応を引き出す手法が確立されれば，実社会における人間を含む系に対するシステム（メカニズム，新規事業サービス，

政策，社会制度など）の機能検証が，生成エージェントを利用した社会シミュレーションにより可能になるだろう．

2. 研究手法

2.1. 生成エージェントの特徴づけ

行動経済学の要素を生成エージェントのペルソナ設定に反映させるため，本研究では Chapman et al. (2023) に焦点を当てた．この研究は意思決定理論をより包括的に理解するための実証的基盤を作る目的で，多数の行動規則性間の相関パターンを調査した．特に，報酬付の調査をアメリカ在住の様々な年齢の 1000 人の参加者に対して行い取得した 21 の行動指標を，主成分分析により 6 つの要素に分割し、これらの関係を分析した。その分析結果を図 1 に示す．本研究ではこの分析結果に焦点を当て，それぞれのクラスター内での重みの内，それぞれで最も大きな値を持つ 6 指標（図 1 でマークアップされている指標）を各クラスターの代表指標として，生成エージェントの特徴づけに利用した．

Chapman et al. (2023) は，分析したパネルデータを公開している．本研究ではそのデータから，各クラスターの 6 つの指標及び CRT，年齢，性別，居住国の各実験参加者からそれぞれ得られた値を用いて，それぞれ異なる 1000 のエージェントを作成した．

	Principal Components						Unexplained Variance
	Generosity	Risk Aversion: WTA	Inequality Aversion/ WTP	Overconfidence	Impulsivity	Uncertainty	
Reciprocity: Low	**.50**	.00	.00	.02	.05	−.01	.29
Reciprocity: High	**.51**	.01	−.02	−.07	.06	.04	.27
Altruism	**.39**	.06	.03	.04	.05	−.03	.54
Trust	**.47**	.01	−.06	.12	−.04	−.05	.33
Antisocial Punishment	−.01	.00	.03	.05	**.67**	−.03	.26
Prosocial Punishment	.09	−.05	−.04	−.08	**.60**	.03	.42
Dislike Having More	.22	.00	**.26**	−.18	−.11	.17	.57
Dislike Having Less	−.06	−.10	**.40**	−.02	.22	.07	.48
WTA	−.04	**−.50**	−.06	.06	.04	.10	.31
Risk Aversion: Gains	.04	**.55**	.11	.03	−.05	.06	.23
Risk Aversion: Losses	−.09	**.36**	−.16	.05	.16	.09	.42
Risk Aversion: Gain/Loss	−.01	**.50**	−.11	−.04	−.01	.03	.24
WTP	−.01	.01	**−.47**	.14	.01	−.02	.45
Risk Aversion: CR Certain	−.02	.10	**.53**	.07	−.02	−.11	.38
Risk Aversion: CR Lottery	−.03	−.03	**.42**	.11	−.03	.04	.56
Ambiguity Aversion	.04	.00	−.03	−.03	.05	**.70**	.25
Compound-Lottery Aversion	−.07	.01	.02	.08	−.07	**.65**	.30
Overestimation	.04	−.03	.01	**.50**	.10	.01	.54
Overplacement	.03	.04	.16	**.49**	−.09	−.05	.56
Overprecision	−.01	−.02	−.07	**.62**	−.02	.06	.33
Patience	.19	−.19	−.05	−.05	**−.28**	.05	.65
Share of variation (%)	14	13	11	8	8	7	40

図 1. Chapman et al. (2023) の主成分分析の結果と焦点を当てた指標

（出所：Chapman et al., 2023; 筆者によりマークアップ）

2.2. シミュレーションの枠組み

本研究では LLM として“gpt-4o-2024-05-13”を利用し，最後通牒ゲームにおける提案側，回答側のそれぞれについて，以下の枠組みでシミュレーションを行った．なお，最後

通牒ゲームは比較対象である Lin et al. (2020) と同様に 100 コインを分配する設定とし，プロンプトは全て英語で記述された．また最後通牒ゲームの説明文は Lin et al. (2020) で用いられたものに倣って作成した．

また，LLM には「temperature」と呼ばれるパラメータがある．GPT においてこれは 0 から 2 の実数値を取り、0 に近いほどより確定的な出力を，2 に近いほどよりランダムな出力を行う．本研究では temperature の設定値として 0, 0.5, 1.0, 1.5 の 4 パターンについて，提案側・応答側それぞれについてシミュレーションを行った．

① パネルデータ（Chapman et al., 2022）を利用し生成エージェントを構成する．

② 各エージェントにゲームをプレイさせる．なおプロンプトの構成は以下の通り．

(ア) 最後通牒ゲームの説明を行う．

(イ) エージェントが意思決定を行う状況の説明を行う．

(ウ) エージェントの出力の形式を指定する．

3. 結果及び考察

得られたシミュレーション結果は図 2 の通りである．まず左図（a）が提案側の結果で，横軸は 100 コインのうちの応答者の取り分のシェアを，縦軸は各値を提案した提案者の割合を表している．青色のヒストグラムは Lin et al. (2020) で用いられた実際の経済実験のデータを表しており，他の色のヒストグラムはそれぞれ異なる temperature の値でのシミュレーション結果を表している．

また右図（b）は応答側について得られたデータをバブルプロットで可視化した結果である．横軸は提案者がオファーした値，縦軸は各提案値に対する受け入れ率を表し，図中の各バブルの中心は各提案値に対する受け入れ率，バブルの大きさはその提案値に対して意思決定を行なったエージェントの数を表している．青色のグラフは Lin et al. (2020) で用いられた実際の経済実験のデータの分析結果，他の色のグラフはそれぞれ異なる temperature の値でのシミュレーション結果を表している．

図 2 (a)，(b)のどちらも，最後通牒ゲームのよく知られる特徴は再現出来ている．提案側については，40~50 の提案値が多く，また相手に半分以上を配分することはほとんどない．応答側については，提案値が 50 に近づくにつれて受け入れ率が上昇していき，50 以上の提案値はほとんど常に受け入れている．さらに，どちらの結果においても temperature の設定値間で結果の差異がほとんどない．この理由としては，エージェントの性格を多角的に設定することにより，想定される反応の幅が狭まっている可能性があると考えられる．

しかし，青色の実際の人間の意思決定と暖色の生成エージェントの意思決定との間にはわずかに差異が存在することが分かる．提案側については，ベンチマークのデータの分布

は 50 にピークが存在するのに対し，シミュレーション結果のピークは 40 である．また応答側については，ベンチマークと比較して，20 以下の提案値についてはより低く，それより高い提案値についてはより高い受け入れ率が示された．

ベンチマークとシミュレーション結果の乖離を埋めるための今後の展開としては，次の２つが考えられる．一つ目は，より質の高いデータを利用することである．ベンチマークの Lin et al. (2020)のデータは様々な大学の授業内で実施された実験の結果であり，十分に統制の取れた環境下で取得されたデータであるとは言い難い．加えて，エージェントの特徴づけに用いた公開データのパネルとベンチマークデータのパネルは異なるため，これらを評価するのは適当でない．よって，同一のパネルから様々な指標と最後通牒ゲームの指標を測定し，そのデータを使った検証を行うことにより，シミュレーション結果と実際の実験結果との適切な比較が実現される．

二つ目は，エージェントの特徴づけに用いる指標を拡充することである．今回利用した指標は十分網羅的でなく，それにより乖離が生じた可能性がある．より多様な指標を含めたパネルデータを構築して生成エージェントの特徴づけにを行うことにより，生成エージェントの反応を実際の人間のものにより近づけることが出来るかもしれない．

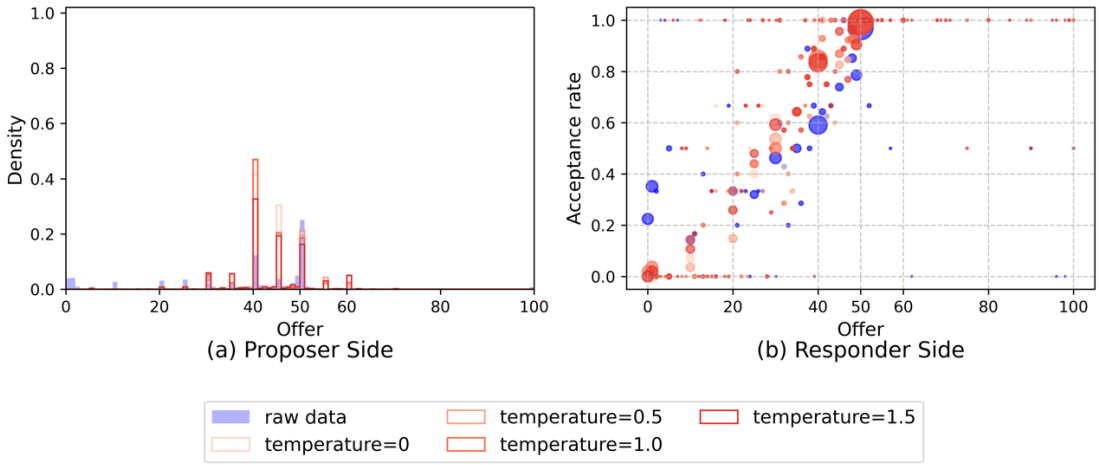

図 2. シミュレーション結果

4. おわりに

本研究では，行動経済学における様々な指標を用いて特徴づけたエージェントによるシミュレーションにより，最後通牒ゲームの実験結果の再現に取り組んだ．結果として，最後通牒ゲームの実験結果としてよく知られる特徴は再現されたものの，ベンチマークとはわずかに差異が存在した．しかし本研究は，LLM に内在する認知バイアスを実際の人間から測定したパネルデータで補正するという新たなアプローチにより，生成エージェント

の出力を実際の人間の反応に近付けられるという大きな可能性を示したと言えるだろう.

引用文献

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.

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.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C., 2022. "Replication Data for: 'Econographics.'" https://doi.org/https://doi.org/10.7910/DVN/IGVOFO.

Chapman, J., Dean, M., Ortoleva, P., Snowberg, E., & Camerer, C., 2023. Econographics. Journal of Political Economy Microeconomics, 1(1), 115-161.

Hagendorff, T., 2023. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:2303.13988.

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.

Kitadai, A., Lugo, S. D. R., Tsurusaki, Y., Fukasawa, Y., & Nishino, N., 2024. Can AI with high reasoning ability replicate human-like decision making in economic experiments?. arXiv preprint arXiv:2406.11426.

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.

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.

Lin, P. H., Brown, A. L., Imai, T., Wang, J. T. Y., Wang, S. W., & Camerer, C. F., 2020. Evidence of general economic principles of bargaining and trade from 2,000 classroom experiments. Nature Human Behaviour, 4(9), 917-927.