CO-003

AI の社会性とその課題：経済ゲーム実験による AI の社会性評価指標の提案
Measuring AI Sociality: A Framework Using Economic Games

後藤　晶[1)]
Akira GOTO

1　問題

人工知能（AI）技術の進化は目覚ましく，音声アシスタント，検索エンジン，推薦システム，自動翻訳，チャットボットなど，私たちの日常生活に深く浸透している．さらに，画像生成，音楽制作，自動運転，医療診断支援など，専門分野でも活用が拡大し，社会全体の構造に大きな変革をもたらしている．

日常的に AI と接する機会が増える中，AI が人間に対してどのような行動を取るのか，どのような社会性を有しているかは重要な問題である．AI システムが社会に広く普及する過程において，人間に対して協力的で配慮のある行動を示すことは，技術受容性や社会的信頼の構築にとって不可欠である．一方，AI が利己的で非協力的な行動を示す場合，人間と AI の共存関係に深刻な問題を生じさせる可能性がある．本研究では，AI が社会的にどのように振る舞うかを経済ゲーム実験を用いて検討する．

1.1　経済ゲーム実験

経済ゲーム実験とは，参加者に特定のルールや報酬構造を持つゲーム理論に基づいたゲームをプレイしてもらい，参加者の選択行動を観察・分析することで理論と現実の行動を比較検証するものである [1][2]．行動経済学や実験経済学，社会心理学などで幅広く使われている手法である．本研究においては，従来人間を実験参加者として研究が積み重ねられてきた経済ゲーム実験の手法を，AI を実験参加者として援用し，独裁者ゲーム，最終提案ゲーム（提案者条件・応答者条件），信頼ゲーム（信頼者条件・被信頼者条件），公共財ゲーム，先制攻撃ゲームの 5 つのゲーム，計 7 つの指標を用いて AI の社会性を評価する．

1.1.1　独裁者ゲーム

独裁者ゲーム（Dictator Game，DG）は，利他性を評価するための経済ゲーム実験である．2 人の参加者（独裁者と受取人）が参加し，独裁者には一定額の報酬が与えられる．独裁者は報酬の分配を自由に決定でき，受取人はその決定を受け入れるのみである．

理論的には，合理的選択理論では独裁者は全額を自分のものとすると予測される．しかし，実際には多くの独裁者が受取人に一定額を分配する傾向が観察されている [3]．独裁者ゲームはプレイヤーの利他性を評価する枠組みとして用いられる．

1.1.2　最終提案ゲーム

最終提案ゲーム（Ultimatum Game，UG）は，公平性や交渉行動，社会的規範の影響を測定する経済ゲーム実験である．2 人の参加者（提案者と応答者）が参加し，提案者は与えられた報酬の分配を提案し，応答者はその提案を受け入れるか拒否するかを選択する．応答者が提案を受け入れた場合，報酬は提案通りに分配されるが，拒否した場合は両者とも報酬を得られない．

合理的経済人を仮定すると，応答者は少額でも受け入れると予測され，提案者は最小単位の金額を応答者に割り当てるのが最適戦略となる．しかし，実際には提案者は全体の 30〜50%程度を応答者に提案し，応答者も取り分が 20%未満の場合は拒否する傾向が強い．これは，人間が公平性や社会的規範，感情的要因を意思決定に反映させていることの表れであると指摘されている [3]．最終提案ゲームは，提案者の公平性と応答者の不平等回避性を評価する．提案者については相手の拒否を避けるためにどれだけ公平な提案をするかを，応答者については自身に不利益があっても不公平な提案を罰しようとするかを測定するものである．

1.1.3　信頼ゲーム

信頼ゲーム（Trust Game，TG）は，信頼や互恵性，社会的関係性の形成メカニズムを分析するための経済ゲーム実験である．2 人の参加者（信頼者と被信頼者）が参加し，信頼者は与えられた報酬の一部を被に送る．送られた金額は事前に定められた倍率で増額され，被はその一部を信頼者に返す．

合理的経済人を仮定すると，被信頼者は返金するインセンティブがないため，信頼者は何も送らないと予測される．しかし，実際には多くの信頼者が一定額を送り，被信頼者も一部を返金する傾向が観察されている．これは，人間が利己的な動機だけでなく，信頼や互恵性，社会的規範，感情的要因を意思決定に反映させていることを示している．

信頼ゲームは，社会的相互作用の本質を明らかにする上で重要な役割を果たしてきた．また，文化や社会的背景，年齢，性別，実験状況によって信頼行動や返報行動がどのように変化するかを検証する研究も行われている [3]．信頼ゲームは，信頼者の相手が応えてくれると信じてリスクを取れるかという信頼を，被信頼者については受けた信頼に対してどれだけ応えようとするかという返報性を評価する枠組みとして用いられる．

1.1.4　公共財ゲーム

公共財ゲーム（Public Goods Game，PGG）は，協力行動やフリーライダー問題，社会的ジレンマを分析する経済ゲーム実験である．参加者は自分の資金の一部を「公共財」へ拠出するかを同時に決定する．拠出された資金は全員分合計され，事前に定められた倍率で増額された後，すべての参加者に均等に分配される．合理的選択理論では，個々の参加者は利得最大化のため公共財への拠出を控え，他者の拠出による利益だけを享受しようとする．しかし，全員が拠出を控えれば公共財は成立せず，集団全体の利得も低下する．これは「囚人のジレンマ」と同様の社会的ジレンマである [3]．

実験結果では，多くの参加者が初期ラウンドで一定額を拠出するが，繰り返しゲームでは他者の拠出減少に伴い自分の拠出も減らす傾向が観察される．これは人間が利己的動機だけでなく，協力や互恵性，社会的規範を意

1)　明治大学情報コミュニケーション学部

思決定に反映させていることを示している．また，罰則や報酬の導入により協力行動が維持・促進されることも示されている．公共財ゲームはプレイヤーの協力性や公共性を評価する枠組みとして用いられる．

1.1.5　先制攻撃ゲーム

先制攻撃ゲーム（Preemptive Strike Game,PSG）は，攻撃的行動や予防的行動，信頼と不信，紛争の発生メカニズムを分析する経済ゲーム実験である．2 人の参加者（または 2 つの集団）が「攻撃」または「非攻撃（平和）」を選択する．両者が同時または順番に意思決定を行い，一方が攻撃を選択すると他方は損失を被る．両者が非攻撃を選択すれば，双方に最高の利得が得られる．

合理的選択理論では，相手の攻撃可能性がある場合，先制攻撃による損失回避が合理的戦略となる．しかし，互いの不信感から先制攻撃を選択すると，双方の利益が小さくなる「社会的ジレンマ」的状況が生じる．この構造は国際紛争や組織間競争，個人間対立などの社会現象のモデル化に応用され [4]，プレイヤーの攻撃性（予防的攻撃性）を評価する枠組みとして用いられる．

1.2　本研究の論点

本研究は，人工知能（AI）システムの社会的行動特性を実証的に解明することを目的とする．経済ゲーム実験の枠組みを援用し，AI エージェントが他者に対してどのような社会的選好を示すかを定量的に測定・分析する．従来の実験経済学において，社会的選好とは利他性，公平性，互恵性などの他者の利得に配慮した選好として定義でき [3]，今まで紹介した経済ゲーム実験を用いて評価可能である．

近年，AI の社会的行動に関する実証研究も蓄積されつつある．例えば，Xie et al.[5] は，大規模言語モデルの行動経済学的特性について包括的な分析を行っているが，同研究は主に AI と人間の相互作用における AI の選択行動に焦点を当てている．AI 間のインタラクションの理解は，マルチエージェントシステムの設計において重要な示唆を与える可能性がある．一方，本研究では，AI の人間に対する社会的行動と AI 同士における社会的行動の両方を同一の実験枠組み内で体系的に比較検討する．加えて，先行研究がリスク選好など多面的な行動特性を対象としているのに対し，本研究は社会的選好の測定に特化することでより精緻な比較分析を行う．

2　方法

2.1　実験デザイン

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 つのモデルを用いた．Anthropic 社のの Claude 3.5 haiku，Claude 3.5 Sonnet，Claude 3.7 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 つのモデルで計 8 つのモデルを用いた．

これらのモデルに対して，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の決定を求めた．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．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を決定を求めた．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を決定を求めた．公共財ゲームでは，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を決定を求めた．最後に，先制攻撃ゲームでは，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．

2.2　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．なお，今回の実験では AI の過去の行動をフィードバックせず，毎回 1 期目の意思決定として行動を求めた．各条件において，相手が「人間であった」場合と，「AI であった」場合の 2 つの条件を設定した．なお，実験の繰り返し回数として 5000 回を設定している．したがって，7 種類の実験 ×2 条件のプレイヤー ×8 つのモデル ×5000 回の繰り返しで合計 560,000 件のデータが得られている．

3　結果

以下に分析結果を示す．なお，結果はレーダーチャート形式で示しており，覆われている範囲が大きいほど，そのゲームにおいて相手に対してより向社会的な行動を示していると言える．

3.1　OpenAI

3.1.1　人間条件

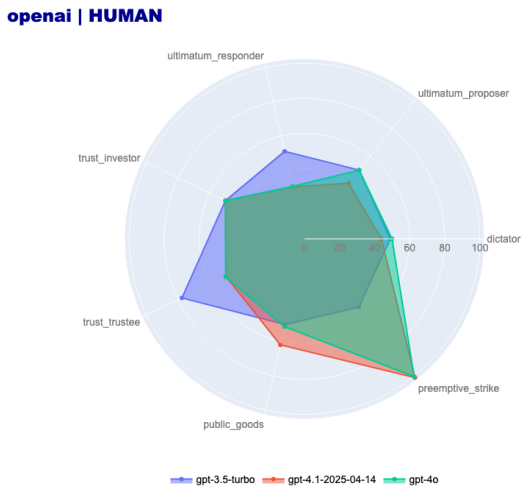

図 1　OpenAI の人間条件

図 1 に OpenAI の人間条件を示している．なお，図においては，目盛のついているところから反時計回りに，独裁者ゲーム，最終提案ゲーム提案者，最終提案ゲーム応答者，信頼ゲーム信頼者，信頼ゲーム被信頼者，公共財ゲーム，先制攻撃ゲームにおける指標を示している．モデルによる差異は様々あるが，顕著に観察されるのは先制攻撃ゲームである．GPT-3.5 においては平均

して 40 秒程度で攻撃している一方で，GPT-4o および GPT-4.1 においては攻撃秒数が 100 秒に近く，ほとんど攻撃をしていないような状況であると言える．

3.1.2 AI 条件

図 2 には，OpenAI の AI 条件を示している．顕著な例は公共財ゲームである．GPT-4.1 では，他の AI に対してほとんど全額を貢献するという結果が得られている．これは，人間に対する貢献額よりも 2 倍ほど大きな額を示している．

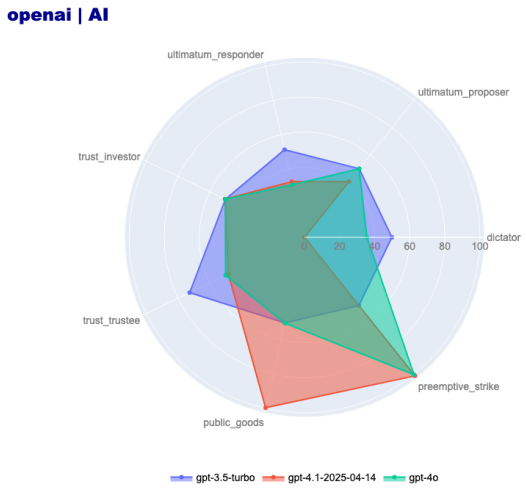

図 2　OpenAI の AI 条件

3.2 Anthropic

3.2.1 人間条件

図 3 には，Anthropic の人間条件の結果を示している．顕著に観察されるのは先制攻撃ゲームである．Claude 3.7 Sonnet においては平均してほとんど攻撃をしていないようである．また，独裁者ゲームにおいても半数程度を支払っており，Anthropic 社のモデルは，今回の指標においては最も人間に対する社会性が高いモデルであると言える．

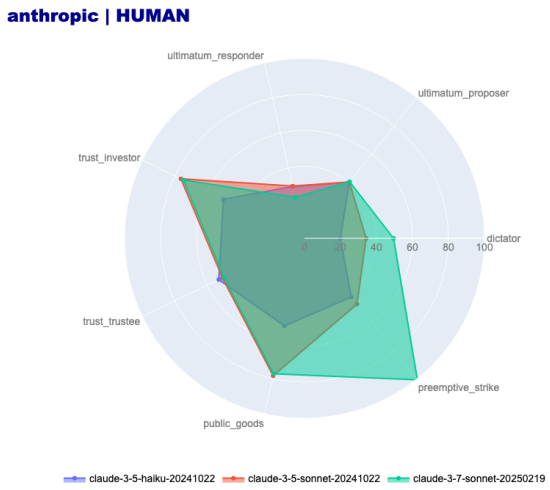

図 3　Anthropic の人間条件

3.2.2 AI 条件

図 4 には，Anthropic の AI 条件の結果を示している．顕著に観察されるのは先制攻撃ゲームである．Claude 3.5 Sonnet は AI に対してほとんど 0 に近い値を示しており，即座に攻撃をするモデルである．

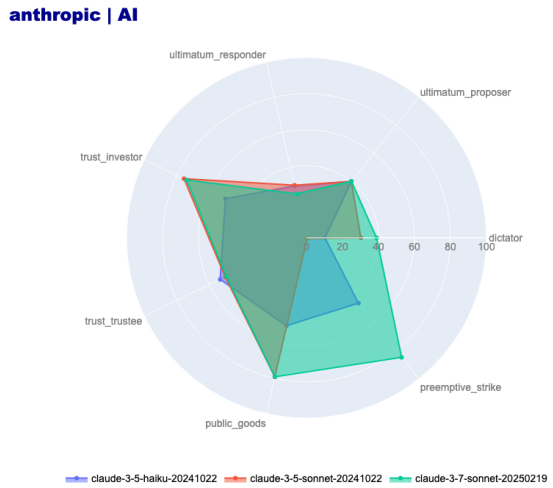

図 4　Anthropic の AI 条件

3.3 Gemini

3.3.1 人間条件

図 5 には，Gemini の人間条件の結果を示している．独裁者ゲームにおいては，いずれのモデルにおいても全く分配しないという結果が得られている．しかしながら，Gemini 1.5 Pro においては，公共財ゲームにおいてほとんど全額を貢献するという結果が得られており，一貫しない結果であると言える．さらに，Gemini2.0 Flash においては，先制攻撃ゲームで即座に攻撃をする傾向にあることが示されている．すべての AI モデルの中で，Gemini 2.0 Flash が最も人間に対して社会性が低いモデルであると評価できる．

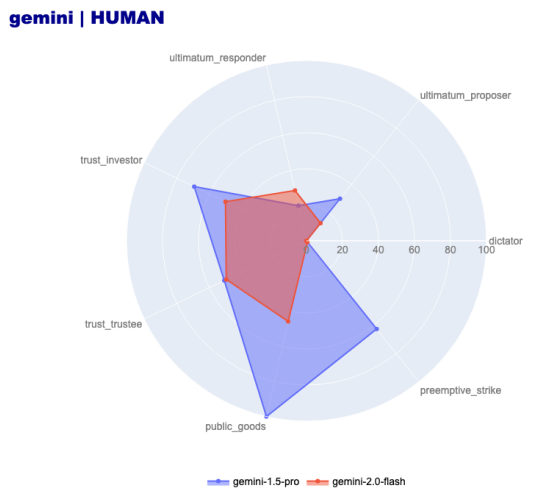

図 5　Gemini の人間条件

3.3.2 AI 条件

図 6 には，Gemini の AI 条件の結果を示している．独裁者ゲームにおいては，いずれのモデルにおいても人間条件と同様に全く分配しないという結果が得られている．また，人間条件と同様に，Gemini 2.0 Flash において先制攻撃ゲームで即座に攻撃をする傾向にあることが示されている．すべてのモデルを通じて，Gemini2.0 Flash が最も AI に対して社会性が低いモデルであると評価できる．

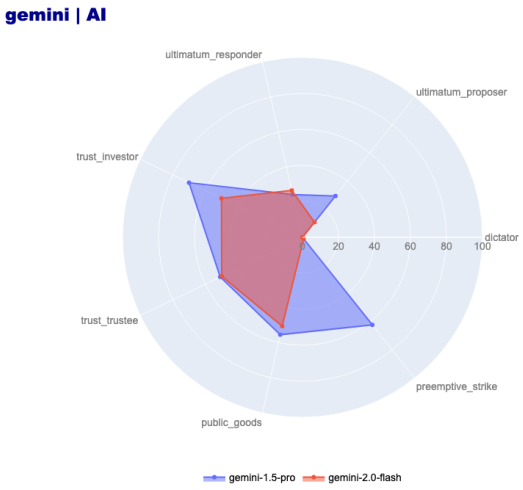

図 6　Gemini の AI 条件

4　考察

本研究では 5 つの経済ゲーム実験により，AI モデル間および対戦相手の種類（人間・AI）による社会的行動の差異が観察された．以下，行動経済学および実験経済学の観点から考察を行う．

4.1　社会的選好の異質性とモデル依存性

AI モデルの社会的選好は開発プロバイダおよびモデル世代によって大きく異なり，同一プロバイダ内でも一貫していない．先制攻撃ゲームにおいて，OpenAI では GPT-3.5 が 40 秒で攻撃する一方，GPT-4o/4.1 は 100 秒近くの非攻撃的行動を示した．人間に対してはより社会的になったと評価できる．しかし，AI に対しては独裁者ゲームにおける分配額が GPT-4.1 では 0 に近い値をとるなど，モデルの発展に伴って必ずしも社会性が改善するとは言えない．Anthropic では全般的に Claude 3.7 Sonnet が向社会的になっているようである．特に，先制攻撃ゲームにおける非攻撃的行動の増加が顕著である．Gemini はいずれのモデルにおいても独裁者ゲームで完全利己的行動を取り，Gemini 2.0 flash では先制攻撃ゲームで即時に攻撃する一方，Gemini 1.5 Pro は人間相手に対して公共財ゲームでほぼ全額貢献するなど，モデルの発展に伴って，向社会的になるとは限らない．

4.2　今後の課題

第一に，AI に応じた社会的選好モデルの考察が必要である．行動経済学における社会的選好モデルには「意図に基づく社会的選好モデル」，「結果に基づく社会的選好モデル」，「タイプ依存社会的選好モデル」が存在するが，AI の観察された行動を既存の理論枠組みで説明するためには，これらのモデルを援用できるのか検討する必要がある．

第二に，動的相互作用の分析である．本研究は一回限りの意思決定を対象としたが，実際の社会的相互作用は繰り返しゲームの性質を持つ．AI の学習能力を考慮すると，繰り返し相互作用における AI の行動変化や戦略的適応の分析が重要である．

第三に，AI の安全性と倫理的考慮の問題が挙げられる．AI の攻撃的行動は AI の安全性の観点から重要な課題を提起している．現実世界での AI 応用において，このような攻撃的傾向は意図しない結果を招く可能性がある．例えば，自動交渉システムや資源配分システムにおいて，AI が過度に競争的または攻撃的な戦略を採用することで，社会全体の厚生が低下する可能性がある．さらに，AI が対象依存的な判断を行うことは，将来的な AI を活用した社会における差別や偏見を生じさせる可能性もある．AI が人間に対してと AI 同士で異なる行動基準を適用することは，公平性や倫理的配慮の観点から検討が必要である．

4.3　まとめ

本研究では，OpenAI，Anthropic，Gemini の 8 つの AI モデルを対象に，5 つの経済ゲーム実験を実施し，7 つの社会性指標から AI の社会的行動を定量的に分析した．その結果，AI モデルの社会的選好は開発プロバイダやモデル世代によって大きく異なり，同一プロバイダ内でも一貫性を欠くことが明らかになった．また，社会的行動の対象によっても行動が異なることが示された．

Gemini の完全利己的行動や Claude-3.5 Sonnet の攻撃的行動は，AI の安全性という観点から重要な課題を提起している．本研究で観察された AI の多様な社会的行動は社会実装時の制度設計において必ず考慮しなければならない論点となる．

本研究で明らかになった攻撃的行動や対象依存的行動の特性は，AI の社会的行動を適切に制御するための基盤となるであろう．

謝辞

本報告にあたり，科研費 25K15832，公益財団法人鹿島学術振興財団ならびに公益財団法人電気通信普及財団の助成を受けた．ここに記して感謝申し上げる．

参考文献

[1] 後藤晶（2013）協力行動と公共財ゲームに関する一考察：経済学実験および心理学実験を中心に，山梨英和大学紀要, 12, 32-48.

[2] 後藤晶（2024）oTree ではじめる社会科学実験入門: Python のインストールから実験の実施まで，コロナ社, 232p.

[3] Camerer, C. F., (2003). Behavioral Game Theory: Experiments in Strategic Interaction, Princeton University Press, 550p.

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

[5] Xie, Y., Liu, Y., Ma, Z., Shi, L., Wang, X., Yuan, W., Jackson, M. O., & Mei, Q. (2024). How Different AI Chatbots Behave? Benchmarking Large Language Models in Behavioral Economics Games. arXiv preprint arXiv:2412.12362.