エージェントとの最後通牒ゲームにおける認知行動モデル －状態遷移確率を用いた検討－

Cognitive-behavioral models in the Ultimatum Game with agents : An investigation using transition matrices

大麻 紀真† 下條 志厳†† 林 勇吾†††

†立命館大学人間科学研究科〒567-8570 大阪府茨木市岩倉町 2−150
††立命館大学グローバル・イノベーション研究機構 〒567-8570 大阪府茨木市岩倉町 2−150
†††立命館大学総合心理学部〒567-8570 大阪府茨木市岩倉町 2−150
E-mail: †cp0150sr@ed.ritsumei.ac.jp

概要

本研究では，異なる行動戦略をとるエージェントとの最後通牒ゲームにおいて，個人の意思決定プロセスを提案行動の選択確率とその遷移確率から検討した．実験の結果，個人は利他的なエージェントに対しては自己の利得を最大化しようとする合理的な戦略をとる一方で，利己的・適応的なエージェントに対しては不公平回避行動をはじめとする経済的合理性から逸脱した反応を示した．この背景として，エージェントの行動戦略に対する予測可能性と，その社会的文脈における解釈可能性の 2 点が重要であることを指摘する．

キーワード:最後通牒ゲーム，意志決定，エージェント，HAI，社会的選好

1. 背景と目的

現代において，人間とコンピュータ上のエージェントとのインタラクションはますます日常的なものとなっており，人間とエージェントの関係においても人間同士の場合と同様に，信頼関係に基づく円滑なコミュニケーションが重要とされている．

より自然で信頼できるエージェントを設計するため，エージェントに対する人間らしさの知覚に注目した研究が数多く行われてきた．この立場における主要なパラダイムは，人間同士のインタラクションと，人間とエージェントとのインタラクションがどのように似ていて，またどのように違っているのかを検討することであった．すなわち，問題は人間同士のインタラクションにおいてみられる意思決定や行動の傾向が，どのような特徴を持つエージェントに対して最もよく表出されるのかというところにある．

人間同士のインタラクションでは，不公平回避行動（Fehr, 1999），すなわち自己の利益を差し置いても不平等な結果を避けようとする傾向や，相手の親切な行動には親切に，不親切な行動には不親切に対応する互恵性・返報性と呼ばれる傾向など，単に経済的な合理性だけでは説明できない社会的選好が観察される．

この点に着目し，社会的選好の古典的な観察課題である最後通牒ゲームを用いてインタラクションの検討が行われてきた．例えば，Hayashi & Okada（2017）は繰り返しのある最後通牒ゲームをエージェントと行う場合，「対戦相手が人間である」という教示によって実験参加者の不公平回避行動が促進されることを示した．また，大津他（2021）では最後通牒ゲームを VR 環境で実施することでエージェントの身体性がその行動意図の理解を助け，互恵的な意思決定を促すと指摘している．加えて，以前の我々の研究でも利己的・利他的・適応的といった異なる戦略をとるエージェントとの最後通牒ゲームを題材に個人の意思決定を検討し，特に利己的なエージェントとのゲームにおいて不公平回避行動が生起しやすい傾向を確認している（大麻他, 2025）．

このように，エージェントが人間らしく社会的な存在として扱われる条件について，教示などによるトップダウン処理，エージェントの身体性や行動戦略によるボトムアップ処理の両観点から分析が行われてきたと言えるが，これらの多くは特定の条件下での行動の結果に着目したものであり，実際にエージェントに対するどのような認知を通して実験参加者の行動が変容していたのかという点については議論の余地が残る．特に，エージェントの戦略に対して個人がどのように行動を適応させていくかという動的なプロセスについて詳細に分析した研究は十分ではない．

そこで本研究では，大麻他（2025）で行った実験における参加者の提案行動についてエージェントの行動戦略との関連から検討を加え，実験参加者の意思決定に関わっていた認知的プロセスを探求する．具体的には，最後通牒ゲームにおける選択行動を，エージェントの応答に対し実験参加者がどのような戦略をもって対応したか，またある行動から次の行動への状態遷移確率

として分析することで，実験参加者の適応的な意思決定についてそのダイナミクスの輪郭を把握する．この研究を通して，エージェントとのインタラクションにおける個人の認知と行動について，より深く理解できることが期待される．

2. 方法

実験には立命館大学の大学生及び大学院生 18 名（男性 15 名，女性 3 名）が参加した．平均年齢は 23.00 歳（*SD*＝3.16）であった．

実験参加者はヘッドマウントディスプレイを装着し，草竹（2020）による装置を用いて VR 空間上のエージェントと最後通牒ゲームを行った． ゲームは実験参加者が報酬の分配方法を表 1 の r1-r7 の中から選んで提案し，エージェントがその提案を承認または拒否する形式で 15 回繰り返して行われた．独立変数としてエージェントは利己的・利他的・適応的のいずれかの戦略をとり，各条件に 6 名の実験参加者が割り当てられた．

各条件のエージェントが実験参加者の提案を拒否する確率を表 1 に示した．利己的エージェントは自己の取り分が少ない提案を拒否する可能性が高く，利他的エージェントは逆に自己の取り分が多い提案を拒否する可能性が高い．適応的エージェントは両極端かつ単純な利己的・利他的エージェントに対し，公平な提案（r4）を承認する一方で，r4, r7 を非対称な確率（30%, 70%）で拒否するなど，バランスを取りながらもより複雑な応答パターンを持つように設計された．

実験参加者による提案の意図を汲みつつ解釈を簡便にするため，提案行動を前回の提案と比べ自身の取り分が多くなるように提案する raise，前回と同じ提案を選ぶ hold，自身の取り分が少なくなるように提案する lower の 3 種類に分類し，それぞれの行動が選択される確率を求めた．

3. 結果と考察

表 2 に実験参加者の前回の提案がエージェントに承認あるいは拒否されている場合にそれぞれの提案行動が選択される確率を示した．また，表 3 には前回・次回ラウンドでの提案行動の遷移確率行列を示している．

先にも述べた通り，最後通牒ゲームにおける人間の意志決定には，不公平回避行動や互恵性・返報性など，経済的な合理性だけでは説明のつかない傾向が存在している．大麻他（2025）での結果と同様，こうした傾向は利己条件において最もよく観察された．例えば，実験参加者は利己的エージェントに対し，前回の自身の提案が拒否されているにもかかわらず raise を選択する確率が最も高かった（0.405）．これは経済的な合理性からは大きく逸脱した，攻撃的ともいえる選択肢だが，このような行動は相手の利己性に対する怒りや反発，Fehr & Gächter（2002）が言うところの「利他的な罰」と捉えることができる．一方で興味深いのは，前回提案が承

表 1　分配方法とエージェントによる拒否確率

	share	probability of rejection		
proposal	participant : agent	ego	exo	adp
r1	**900 : 100**	95%	5%	30%
r2	**800 : 200**	80%	20%	20%
r3	**700 : 300**	80%	20%	10%
r4	**500 : 500**	100%	0%	0%
r5	**300 : 700**	20%	80%	40%
r6	**200 : 800**	20%	80%	70%
r7	**100 : 900**	5%	95%	70%

表 2　前回提案の成否と次回の提案行動の選択確率

		last proposal	
		accepted	rejected
ego	raise	0.190	0.405
	hold	0.429	0.310
	lower	0.381	0.286
exo	raise	0.197	0.391
	hold	0.705	0.087
	lower	0.098	0.522
adaptive	raise	0.182	0.389
	hold	0.576	0.389
	lower	0.242	0.222

表 3　前回・次回ゲームでの提案行動の遷移確率行列

		next		
	previous	raise	hold	lower
ego	raise	0.284	0.307	0.410
	hold	0.356	0.346	0.298
	lower	0.354	0.344	0.302
exo	raise	0.311	0.362	0.327
	hold	0.213	0.487	0.300
	lower	0.392	0.258	0.350
adaptive	raise	0.364	0.374	0.262
	hold	0.403	0.425	0.171
	lower	0.396	0.416	0.188

認された場合に実験参加者が lower を選択する確率が 0.381 と，他の条件と比べてかなり高くなっている点である．これは互恵性に基づく意志決定の可能性があり，また「相手の厚意につけ込むべきではない」といった社会的規範が働いた結果とも考えられる．表3を見ると，実験参加者による提案行動の遷移確率は他の条件と比べて安定した均衡点を持たず，戦略が揺れ動いている様子が見て取れる．例えば，raise の後では lower が選ばれる確率が 0.410 と高く，また逆に lower の後では raise が選ばれる確率が 0.354 と高いことから，実験参加者がより攻撃的な行動と譲歩的な行動を切り替えながら繰り返していたことが分かる．このパターンは互いの利益をなるべく均等にしようとする，不公平回避行動の表れと解釈できる．さらに，他の条件で見られる安定した均衡，すなわち hold への収束を示さない点を踏まえると，ここに感情的な対立や駆け引きが影響していた可能性も否定できない．つまり，実験参加者が強気に出て（raise）拒否され，不満を抱えつつ譲歩し（lower），しかしそこで引き下がらずにまた強気に転じる（raise）といった，エージェントとの絶え間ない綱引きに終始していたという解釈も可能である．いずれにせよ，エージェントが利己的な行動を取る場合に，実験参加者は不公平回避行動をはじめとする，経済的合理性から逸脱した反応を示しやすい可能性が示唆された．

反対に，表 2 から利他条件ではむしろ実験参加者の合理的な意思決定が強く現れていたことが読み取れる．例えば，実験参加者は前回の自身の提案が承認されている場合，0.705 という高い確率で hold を選択し（win-stay），また拒否されている場合には 0.522 とこちらも高い確率で lower を選択する（lose-shift）傾向にあった．これは典型的な「パブロフ」戦略（Nowak & Sigmund, 1992）であり，成功した手段を繰り返し，失敗すれば他の手段を試すことで相手の行動に対応しつつ最大利益を獲得しようとする適応的なプロセスといえる．表 3 の遷移確率を見ると，実験参加者は一度 hold を選択した場合，次回も hold を選択する確率が 0.487 と高くなっている．これは先の win-stay 戦略を裏付ける結果であり，参加者は一度うまくいく提案を見つけると，その戦略を継続する傾向にあることが分かる．

適応条件について，表 2 を見ると前回提案が承認のときに hold を選ぶ win-stay 戦略が見て取れるが，利他条件ほど突出して選ばれやすいわけではない．一方で拒否された場合に raise と hold が同程度に選ばれており，利己条件に近いエージェントへの反発のような反応が見て取れる．提案行動の遷移確率は利他条件に近く，hold から hold への遷移確率が 0.425 と最も高い．

適応的エージェントの提案拒否確率はどちらかといえば利他条件に近く設定されていたため，提案行動の遷移確率が両条件で似通っていることに不思議はなく，これは大麻他 (2025) で示された利己条件に対してのみ不公平回避行動が生起しやすいという実験結果とも矛盾しない．しかし一方で表 2 に示されたエージェントの承認・拒否に対する対応を見ると，適応的エージェントに対する実験参加者の戦略はむしろ利己条件のそれに近い．このことから，実験参加者の行動戦略の変容には，エージェントの利己性だけでない何らかの要因が関与していたことがうかがえる．考えられる要因として，エージェントの行動に対する予測可能性が挙げられる．実験参加者により多くの利益を譲るという利他的エージェントの一貫した戦略は，実験参加者がかれを複雑な意図を持った主体としてではなく，むしろ明確な規則に従って動く「攻略対象のシステム」として認識するように仕向けた可能性がある．この明確さゆえに，参加者は複雑な社会的駆け引きを考慮する必要がなく，システムのルールを効率的に学習し選択を最適化する合理的な行動パターンをとっていたのではないだろうか．利己条件も利他条件と同様に予測可能性の高いシステムであったが，参加者より多くの利益を獲得しようとする戦略は実験参加者にとって不当な搾取，あるいは対話の拒絶といった社会的な意味合いを帯びていたため，「攻略対象」ではなく「罰するべき理不尽な相手」という解釈が優先して適用され，結果として社会的な反応を誘発したのだと考えられる．そして，適応条件ではこれらの 2 条件の間を取ったような拒否確率が設定されていたことで，実験参加者がその戦略を予測することが難しくなり，結果として「複雑な意図を持った対戦相手」としての評価が高まったのだろう．

しかし，それでいて適応条件が利己条件ほど社会的あるいは感情的な反応を誘発しなかったということは，予測可能性が対戦相手の捉え方に対し単に一方向に作用するものでないことを示唆している．期待違反理論（Burgoon, 1993）によれば，人間はコミュニケーションの相手に対し社会的規範や既有の知識と照らし合わせて妥当な行動パターンを期待する傾向にあり，そこから逸脱した行動をとる相手とのコミュニケーションはネガティブに評価されやすい．大麻他（2025）で示した通り，最後通牒ゲームにおける人間の意志決定は基本的に相手が譲歩する限り自己の利益を増やそうとする

利己的なものであるから，エージェントの極端な譲歩は実験参加者にとって理解しにくいものだった可能性がある．結果として利他的エージェントは社会的な文脈を理解しておらず，積極的にコミュニケーションをとるに値しない相手と評価されたのかもしれない．逆に予測可能性の高い戦略であっても，利己的エージェントの「自己の有利を追求する」のように自然かつ理解可能な方針に従うものであれば，かえって主体性の認知を高め，人間らしい社会的なコミュニケーションを促進する可能性があるといえるだろう．

4. まとめ

本研究では，エージェントの行動戦略（利己的・利他的・適応的）が最後通牒ゲームにおける個人の意思決定プロセスに与える影響を，提案行動の選択傾向とその遷移確率から検討した．その結果，実験参加者の提案行動はエージェントの戦略によって質的に大きく異なることが明らかになった．利他的エージェントに対しては，実験参加者は自己の利得を最大化しようとする合理的な戦略を示した．対照的に，利己的エージェントに対しては提案が拒否された後にあえて要求を引き上げるなど，経済的な合理性から逸脱した感情的・社会的な反応が見られた．適応的エージェントに対しては安定した妥協点を探し，維持しようとする戦略をとるという点では利他条件と共通していたものの，利己条件と似た非合理な意思決定を行う様子も観察された．

これらの差を説明する鍵として，本研究ではエージェントの行動戦略に対する予測可能性と，その内容の社会的文脈における解釈可能性という 2 点の重要性を指摘する．利他的エージェントはその戦略が一貫しており，かつ社会的な文脈にそぐわないため，実験参加者に「攻略対象のシステム」として認識され，合理的な選択行動を促進したのではないかと考えた．一方で，利己的エージェントの戦略は予測可能であってもその内容が「不当な搾取」といった社会的な意味合いを帯びたため，参加者の社会的な反応を誘発したのだと考えた．

本研究を通して，エージェントの行動戦略は個人がその利得を最大化しようとする学習プロセスだけでなく，相手の意図を推測し，それに対して感情的に反応するという，より複雑な認知プロセスを誘発することによってその行動を変容させている可能性が示唆された．しかしながら，実験参加者のエージェントに対する主観的評価については直接の検証を行ったわけではなく，あくまでその行動から推察するにとどまっている．今後は質問紙調査によって，エージェントが実際に実験参加者からどのように捉えられていたのかという点についても検討を加えていきたい．並行して，これまでに得た知見から不公平回避などの社会的選好に加え，怒りや遠慮といった感情的な反応，相手への期待や予測といったパラメタを組み込んだ認知プロセスのモデリングも行っていく．

本研究で得られた知見は，人間と協調する AI アシスタントや対話エージェントの設計に重要な示唆を与える．利己的なエージェントは人間の感情的な反発を招くおそれがある．しかし，社会的文脈から大きく逸脱した「利他的すぎる」エージェントもまた，かえって社会的な反応を阻害する．人間らしい感情や駆け引きを含む円滑で豊かなインタラクションを可能にするためには，エージェントの設計に際し人間がその行動をどう解釈し，どのような意味付けを行うかという特性を考慮に入れることが不可欠といえるだろう．

文献

Burgoon, J. K. (1993). Interpersonal expectations, expectancy violations, and emotional communication. *Journal of language and social psychology*, *12*(1-2), 30-48. http://dx.doi.org/10.1177/0261927X93121003

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137-140. https://doi.org/10.1038/415137a

Fehr, E., & Schmidt, K., M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817-868. https://doi.org/10.1162/003355399556151

Hayashi, Y., & Okada, R. (2017). Compound effects of expectations and actual behaviors in human-agent interaction: Experimental investigation using the Ultimatum Game. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 2168-2173. https://escholarship.org/uc/item/72t3c633

草竹 大輔（2020）. VR 環境における最後通告ゲームを用いた実験的検討　立命館大学卒業論文（未公刊）

Nowak, M. A., & Sigmund, K. (1992). Tit for tat in heterogeneous populations. *Nature*, *355*(6357), 250-253. http://dx.doi.org/10.1038/355250a0

大麻 紀真・下條 志厳・林 勇吾・渡邊 咲花 (2025). エージェントの行動戦略が個人の意思決定に与える影響――最後通牒ゲームを用いた実験的検討　信学技報, 124, 188-193.

大津 耕陽・林 勇吾・下條 志厳・田村 昌彦・泉 朋子 (2021). VR 環境における不公平回避行動に関する分析 2021 年度日本認知科学会第 38 回大会発表論文集, 205-211.