Vol.2020-IS-154 No.7
2020/12/12

人は人工知能と協力できるのか：公共財ゲームを用いた予備的検討

後藤　晶 [1]

概要：現代社会において，人工知能の開発が着々と進められている．技術的な開発は様々に行われつつあり，例えばAlpha Go は強化学習を用いたコンピュータエージェントとして世界一のプロの棋士との「勝負」にも勝てる能力を有するまで至った．一方で，人間が人工知能との「協力」できるかどうかはもう一つの論点になりえるであろう．本報告においては，公共財ゲームの枠組みを用いて，クラウドソーシングを利用することにより，実験的にあたかも自律的であるかと思えるような反応を返す人工知能と人間の協力行動の可能性について検討する．

キーワード：協力行動，公共財ゲーム，人工知能，クラウドソーシング

Can People Cooperate with Artificial Intelligence: A Preliminary Analysis Using Public Goods Games

AKIRA GOTO[†1]

Keywords: Cooperative Behavior, Public Goods Game, Artificial Intelligence, Crowdsourcing

1. はじめに

言うまでもなく，昨今では人工知能の発展が著しい．例えば，Google DeepMind により開発された Alpha Go は 2017 年 5 月に世界トップ棋士である柯潔との三番勝負で全勝するなど[1]，人間を圧倒する能力を有しつつある．

一方で，これからの社会には単純に人工知能と「競争」したり，人工知能を「使う」だけでは済まされない時代が来るものと考えられる．特に，今重要視されているのは人間と人工知能との「コラボレーション」である．

その中で，一つの論点は人工知能と人間は「協働」「協力」できるかであろう．Crandall らは囚人のジレンマを用いた実験によって，人間と人工知能エージェント同士の協力が，人間同士の協力より少ないこと，人間と人工知能エージェントの間に選択式のチープトークを可能にすると人間と人工知能エージェントの協力行動が，人間同士の協力と同程度に観察されたことが報告している[2].

これらの観点を踏まえて，本研究の目的は人工知能と「協力」しやすい人の特徴について解明することにある．協力行動を分析する枠組みの一つである公共財ゲームを用いて，「利己的な人工知能」「中立的な人工知能」「協力的な人工知能」「ランダムな人工知能」の 4 種類の人工知能との協力行動について検討する．

2. 方法

本研究では「Yahoo! クラウドソーシング (http://crowdsourcing.yahoo.co.jp/)」を用いた．調査は 2020 年 10 月 30 日 17 時 00 分から 23 時 25 分にかけて実施した．調査参加者は 1,197 名(年齢 M=44.825，SD=10.65，年齢回答を拒否した方を除く)，内訳は男性が 742 名(年齢 M=46.59, SD=9.92)，女性が 847 名(年齢 M=41.64，SD=11.07)，性別を回答しない方が 7 名(年齢 M=37,71, SD=16.98)であった．

今回実施した公共財ゲームの利得関数は以下の通りである．グループ j に所属するプレイヤーi の利得π_{ij}は，C_iをプレイヤーi の貢献額，$\sum C_j$をグループ j におけるプレイヤーの貢献額の合計とすると，$\pi_{ij} = 100 - C_i + 2/3\sum C_j$として表すことができる．初期保有額を 100 ポイントとして，3 人プレイヤーで実施している．実験時には 3 人プレイヤーのうち 2 体を人工知能エージェントとして設定している．人工知能エージェントについては，0-33 ポイントをランダムに貢献する Selfish 群，34-66 ポイントをランダムに貢献する Neutral 群，67-100 ポイントをランダムに貢献する Cooperative 群としており，これらの 4 群の組み合わせとして 16 群について実験を行った．

さらに，公共財ゲームに加えて，同ゲームの理解度確認問題，一般的信頼尺度[3]，SVO スライダー[4]，認知反射テスト[5]，および性別・年齢・居住地域等の社会経済的要因に関する項目であった．

なお，一般的信頼尺度および SVO スライダーについては，従来の「対人」の尺度に加えて，「対人工知能」に変更した尺度についても調査を行った．これらの実験・調査システムは oTree を用いて開発した[6].

1 明治大学　情報コミュニケーション学部
Meiji University, School of Information and Communication.

ⓒ2020 Information Processing Society of Japan

3. 結果

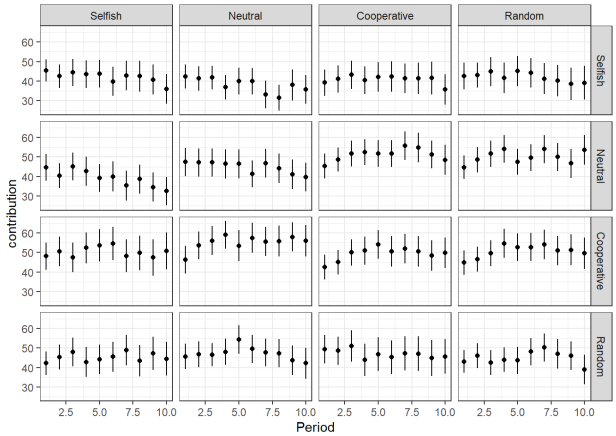

図 1 各条件における平均貢献額

図 1 には各条件における平均貢献額のプロットを示しており，表 2 には記述統計量を示している．

分析結果を表 2 に示す．Model 1 は繰り返しの期数，チェック問題の正答数に加えて，対 AI 一般的信頼尺度，対人一般的信頼尺度，対 AI 社会的価値志向性（SVO），対人社会的価値志向性ならびに認知反射テストのスコアを投入したものである．このモデルからは，期を経る毎に貢献額が減少すること，チェック問題の正答数が多く，公共財ゲームのルールを理解している人ほど貢献額が減少していること，さらに認知反射テストのスコアが高いほど貢献額が減少することが示されている．特に，対人社会的価値志向性は人対人の公共財ゲームにおいてはポジティブな影響が示されるものである．しかしながら，今回の対人工知能条件においては，対人社会的価値志向性は影響が認められず，対 AI 社会的価値志向性がポジティブな影響を示されることとなった．

続いて，Model 2 は Model 1 から対人一般的信頼ならびに対社会的価値志向性を除いたモデルである．このモデルにおいては Model 1 と比較すると認知反射テストの影響が認められなくなっている．

さらに，Model 3 は Model 2 に加えて，各実験条件を加えたものである．実験条件についてはいずれも有意差が認められなかった．

この 3 つのモデルについて AIC を基準に評価を行ったところ，Model 2 の AIC が最小の値であった．したがって，今回は Model 2 を最良のモデルとして評価する．

4. ディスカッション

本研究の結果をまとめると，以下の通りである．

- 期を経るほどに貢献額は減少する．
- チェック問題の正答数が多いほど，貢献額は減少する．
- 対人一般的信頼および対人社会的価値志向性は影響を及ぼさない．
- 今回の分析では実験条件は協力行動に影響を及ぼさない．

	Overall (N=11970)
貢献額	
Mean (SD)	46.0 (32.6)
Median [Min, Max]	50.0 [0, 100]
Period	
Mean (SD)	5.50 (2.87)
Median [Min, Max]	5.50 [1.00, 10.0]
	Overall (N=1197)
条件	
Cooperative-Cooperative	77 (6.4%)
Cooperative-Neutral	79 (6.6%)
Cooperative-Random	70 (5.8%)
Cooperative-Selfish	68 (5.7%)
Neutral-Cooperative	65 (5.4%)
Neutral-Neutral	79 (6.6%)
Neutral-Random	69 (5.8%)
Neutral-Selfish	82 (6.9%)
Random-Cooperative	76 (6.3%)
Random-Neutral	77 (6.4%)
Random-Random	86 (7.2%)
Random-Selfish	71 (5.9%)
Selfish-Cooperative	61 (5.1%)
Selfish-Neutral	77 (6.4%)
Selfish-Random	76 (6.3%)
Selfish-Selfish	84 (7.0%)
チェック問題正答数	
Mean (SD)	5.18 (1.94)
Median [Min, Max]	5.00 [0, 8.00]
対AI一般的信頼	
Mean (SD)	19.8 (4.98)
Median [Min, Max]	20.0 [5.00, 35.0]
対人一般的信頼	
Mean (SD)	18.8 (5.64)
Median [Min, Max]	20.0 [5.00, 35.0]
対AISVO	
Mean (SD)	22.3 (11.7)
Median [Min, Max]	22.6 [-16.3, 61.4]
対人SVO	
Mean (SD)	24.4 (11.7)
Median [Min, Max]	22.8 [-16.3, 61.4]
認知反射テスト	
Mean (SD)	1.52 (1.15)
Median [Min, Max]	2.00 [0, 3.00]

表 1　記述統計量

ⓒ2020 Information Processing Society of Japan

	貢献割合					
	Model 1		Model 2		Model 3	
Predictors	*Odds Ratios*	*p*	*Odds Ratios*	*p*	*Odds Ratios*	*p*
(Intercept)	0.308 ** (0.126 – 0.753)	**0.01**	0.369 * (0.158 – 0.862)	**0.021**	0.331 * (0.131 – 0.837)	**0.019**
Period	0.992 *** (0.990 – 0.993)	**<0.001**	0.992 *** (0.990 – 0.993)	**<0.001**	0.992 *** (0.990 – 0.993)	**<0.001**
チェック問題正答数	0.833 *** (0.775 – 0.895)	**<0.001**	0.833 *** (0.776 – 0.895)	**<0.001**	0.832 *** (0.775 – 0.893)	**<0.001**
対AI一般的信頼	1.014 (0.986 – 1.042)	0.345	1.02 (0.995 – 1.046)	0.123	1.015 (0.989 – 1.041)	0.259
対人一般的信頼	1.014 (0.989 – 1.040)	0.27				
対AISVO	1.032 *** (1.018 – 1.046)	**<0.001**	1.035 *** (1.024 – 1.046)	**<0.001**	1.035 *** (1.024 – 1.046)	**<0.001**
対人SVO	1.005 (0.991 – 1.019)	0.518				
認知反射テスト	0.880 * (0.778 – 0.994)	**0.04**	0.887 (0.786 – 1.002)	0.055	0.891 (0.789 – 1.006)	0.062
Player2：（ctrl：Selfish）						
Neutral					0.985 (0.511 – 1.898)	0.963
Cooperative					1.088 (0.543 – 2.179)	0.812
Random					0.825 (0.415 – 1.643)	0.585
Player3：（ctrl：Selfish）						
Neutral					0.839 (0.429 – 1.642)	0.608
Cooperative					1.801 (0.883 – 3.673)	0.106
Random					0.962 (0.492 – 1.885)	0.911
交互作用						
Neutral * Neutral					1.499 (0.580 – 3.870)	0.403
Cooperative * Neutral					2.078 (0.787 – 5.491)	0.14
Random * Neutral					2.156 (0.819 – 5.673)	0.12
Neutral * Cooperative					1.287 (0.474 – 3.491)	0.621
Cooperative * Cooperative					0.624 (0.229 – 1.703)	0.357
Random * Cooperative					1.369 (0.502 – 3.732)	0.54
Neutral * Random					1.317 (0.503 – 3.450)	0.575
Cooperative * Random					1.016 (0.377 – 2.737)	0.975
Random * Random					1.091 (0.417 – 2.853)	0.858
社会経済的要因（性別・年齢・居住地域・年収・未既婚・子の有無）を統制済み						
Random Effects						
σ2	3.29		3.29		3.29	
τ00	4.61 code		4.62 code		4.51 code	
ICC	0.58		0.58		0.58	
N	1197 code		1197 code		1197 code	
Observations	11970		11970		11970	
Marginal R2 / Conditional R2	0.060 / 0.608		0.059 / 0.608		0.070 / 0.608	
AIC	347623.488		347621.3		347626.265	

表 2　分析結果

チェック問題正答数がネガティブな影響を与えていることは，公共財ゲームに関する理解度が高いほど利己的であることを示している．

社会的価値志向性については対人社会的価値志向性の影響は認められず，対 AI 社会的価値志向性のポジティブな影響が認められた．従来，人間同士で公共財ゲーム実験を実施する際には対人社会的価値志向性の影響が認められていたが，協力行動の対象が異なると，影響を与える社会的価値志向性も異なることが確認できた．

また，今回の分析からは実験条件がいずれも影響を及ぼすことはなかった．この結果は，条件付き協力の観点を考慮すると[7]，協力行動を行うエージェントに対して，人は協力するはずである．しかしながら，今回はその傾向が観察されなかった．本研究においては人工知能がどのように振る舞おうと人間の行動は変化せず，人工知能に対する態度は個人の資質のみによるところが大きいことが示唆されている．

ただし，今回条件を考慮した Model 3 は条件別で分析しているものである．実際の貢献額に基づいた分析など，さらなる分析検討が必要になるために，本研究の結果はあくまでも限定的な結果である．

謝辞　本研究は JSPS 科研費 19K20634 の助成により実施しました．ここに記して感謝申し上げます．

参考文献

[1] Wikipedia, Alpha Go, https://ja.wikipedia.org/wiki/AlphaGo, 2020 年 11 月 13 日閲覧
[2] Crandall, J.W et.al, Cooperating with Machines, Nature Communications, vol.9, no.233, 2018
[3] Yamagishi, T. et al, Is behavioral pro-sociality game-specific? Pro-social preference and expectations of pro-sociality. Organizational Behavior and Human Decision Processes, vol.120, no.2, pp.260–271.
[4]Murphy, R. O., et.al: Measuring Social Value Orientation (SVO). Judgment and Decision Making, vol.6, pp.771-781, 2011.
[5]Frederick,S.: Cognitive reflection and decision making, Journal of economic perspectives, vol.19, pp.25-42, 2005.
[6]Chen, D.L, et.al, :oTree—An open-source platform for laboratory, online, and field experiments, Journal of Behavioral and Experimental Finance, vol.9, pp.88-97, 2016
[7]Fischbacher U et.al, :Are people conditionally cooperative? Evidence from a public goods experiment. Economic Letters vol.71, no.3, pp.397–404, 2011.

ⓒ2020 Information Processing Society of Japan