善人と悪人を識別する認知計算
Cognitive computation for distinguishing good from bad

寺田 和憲[†]，長谷川 智大[†]，セルス ドゥメル[‡]，フランシスコ サントス[¶]
Kazunori Terada[†], Tomohiro Hasegawa[†], Celso M. de Melo[‡], Francisco C. Santos[¶]
[†] 岐阜大学，[‡]DEVCOM ARL，[¶] リスボン大学
[†]Gifu University, [‡]DEVCOM ARL, [¶]Universidade de Lisboa
kazunori.terada@acm.org

概要

協力がより高い利益を生む一方で，協力的な行為が搾取される可能性のある非ゼロ和的社会においては，搾取者（悪人）を避けながら協力者（善人）と良好な関係を構築することが重要な課題である．しかし，インタラクションの中で，どのような認知計算によって，人が善人と悪人を識別しているかについては明らかになっていない．本研究では，人が，自他の相対利益を評価をする性格特性のモデルを用いて，観察した相手の行動から相手の性格の善悪をベイズ推論し，新規状況で相手の行動の善悪を予測し，合理的な意思決定ができるかどうかを実験によって確かめた．実験参加者（$n = 372$）は，競争的から協力的までの 4 段階のいずれかの性格特性を持ち，性格特性によって異なる意思決定と表情表出をする AI エージェントと，10 ラウンドの鹿狩りゲームに続く 5 ラウンドの囚人のジレンマをプレイした．実験の結果，実験参加者は 10 ラウンドの鹿狩りゲームにおいて，AI エージェントの振舞いから AI エージェントの性格特性を推論し，新規ゲームである囚人のジレンマにおいて合理的な意思決定をすることがわかった．この結果は，人が自他の利益に関する評価モデルを用いて他者の善悪を推論し，意思決定に活かしていることを示唆する．

キーワード：善悪，心の理論，社会的価値志向性，逆評価

1. はじめに

利他行動は短期的には自分の利得を減らし，相手に与える一見不合理な行動であるが，長期的には他者の協力や返報を受けることで，利得の向上が図れるために合理的である場合もある [1]．しかし，利他行動は悪人による搾取の被害にあう可能性があるために，利益を与える相手は善人（協力者もしくは利他者）に限定しなければならない [2]．見知らぬ相手が善人か悪人であるかを知るために，相手の過去の振る舞いに対して誰かが善悪のラベルを付けした「評判」が有効であることが知られている [3] が，直接的に相手の性格や意図などの心的状態を推測することができれば，より正確に相手が協力的か搾取的であるかを見分けることができると考えられる．

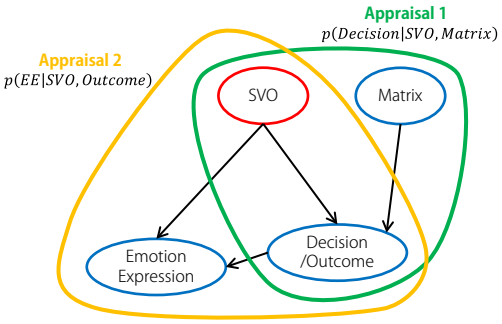


図 1: 同時手番ゲームにおける 2 つの評価過程（尤度）．Appraisal 1 は性格特性と利得表によって選択を決定する評価過程．Appraisal 2 は性格特性とゲーム結果を与えると表情表出を決定する評価過程．

相手の性格特性の推測は，相手が状況をどのように評価するかについての生成モデルを尤度として用いて，観察した相手の行動を入力としてベイズ推論を行うことで行われると考えられる [4, 5]．例えば，協力的な性格特性の人は囚人のジレンマで協力を選択し，個人主義的な性格特性の人は囚人のジレンマで裏切りを選択するだろう，という一般的知識を人は持っていると考えられる．この知識を用いて，裏切りを選択した相手の性格が個人主義的な性格である可能性が高いという推論ができる．また，相互協力という両者の意思決定結果をポジティブに評価する（喜びを表出する）相手は協力的性格である可能性が高いと推論できる．この推論は図 1に示す 2 つの評価過程をそれぞれ尤度として用いた次の 2 つのベイズ推論によって実現される．

$$p(SVO|M, D) \propto p(D|M, SVO)p(SVO) \quad (1)$$

$$p(SVO|O, EE) \propto p(EE|O, SVO)p(SVO) \quad (2)$$

ここで，SVO は性格特性（社会的価値志向性，後に詳述する），M は利得表，D は意思決定，EE は表情表出，O は開示された両者の意思決定結果である．式

2において式 1の事後確率を事前確率として用いることで SVO の確信度の更新ができる．また，反復ゲームの場合は一つ前のラウンドの事後確率を事前確率として用いて確信度の更新ができる．

人が合理的エージェントの生成モデルを用いて観察した相手の行動 [4] や表出表情 [5] から相手の心的状態の推論と行動予測することについて理論的な検討 [6] がなされ，認知計算モデルの提案と実証が行われてきたが，社会的関係の概念の上位に位置する善悪の性格特性を人が他者に帰属させて行動予測をするかについては検証されていない．本研究では，実験参加者に，競争的から協力的までの 4 段階のいずれかの性格特性を持ち，性格特性によって異なる意思決定と表情表出をする AI エージェントと，10 ラウンドの鹿狩りゲームに続く 5 ラウンド囚人のジレンマを反復プレイすることを求め，調整ゲームである鹿狩りゲームの中で同定した相手の善悪を新規の混合動機ゲームである囚人のジレンマでの意思決定に活かせるかどうかを調べた．なお，異なるゲーム間での一貫した戦略は連結と呼ばれる [7] が，善悪の性格特性を媒介とした連結についてはこれまで未検討である．

2. 実験

2.1 実験参加者と実験計画

実験参加者は Yahoo!クラウドソーシングで募集した 20 歳から 70 歳までの男性 257 人，女性 111 人，そのほか 4 人であった（$M_{age} = 46.71, SD_{age} = 10.62$）．実験計画は 2（表情表出: あり・なし）× 4（SVO の角度:45 度・5 度・-5 度・-45 度）の参加者間要因であった．

2.2 善悪と合理的意思決定

我々は社会的価値志向性（social value orientation : SVO）を善悪の性格特性を特徴づける評価モデル（appraisal model）として用いた．SVO とは，お金などの資源を分配する際に自他にどのような重み付けをして配分するかの好みであり [2]（図 2a参照），一見非合理であるが長期的に合理的な利他行動を熟考なく生成する動機となる．SVO と利得表を与えると合理的意思決定は一意に決定する（表 2c参照）．すなわち式 1の推論に用いる尤度 $p(D|M, SVO)$ を生成できる．エージェントがある選択をした場合の期待報酬 U は

$$U = [w_{self}, w_{other}] \cdot [R_{self}, R_{other}], \tag{3}$$

としてあらわすことができる．ここで，w_{self}，w_{other} は SVO によって決定される，自他の価値を重視する程度をあらわす重みである．R_{self}，R_{other} は当該選択が行われた場合の自他それぞれの報酬である．表 2cの鹿狩りゲームの 2 つの選択 S と H について，それぞれの期待報酬 U_S，U_H は式 4，5で計算でき，その結果を用いて合理的意思決定は式 6，7で計算できる．

$$U_S = [p_{\mathrm{S'}} \cdot \mathrm{SS'_u} + p_{\mathrm{H'}} \cdot \mathrm{SH'_u}] \tag{4}$$

$$U_H = [p_{\mathrm{S'}} \cdot \mathrm{HS'_u} + p_{\mathrm{H'}} \cdot \mathrm{HH'_u}] \tag{5}$$

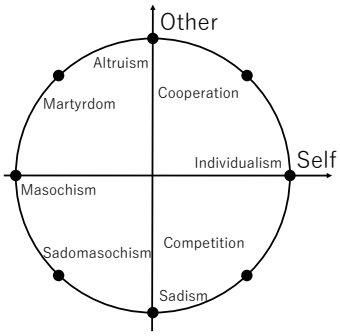
$$Rc = U_{\mathrm{S}} - U_{\mathrm{H}} \tag{6}$$

$$\text{Agent Choice} = \begin{cases} \mathrm{S} & \text{if } Rc > 0, \\ \mathrm{Random} & \text{if } Rc = 0, \\ \mathrm{H} & \text{if } Rc < 0, \end{cases} \tag{7}$$

協力ゲームである鹿狩りゲームのバリエーションは図 2bに示す 5 種類であるが，異なる SVO が異なる意思決定を出力するように SVO 角度を 45（協力的）, 5（協力寄りの個人主義）, -5（競争寄りの個人主義）, -45 度（競争的）に決定した．5 種類の鹿狩りゲームに対する協力（S）の選択確率は 45 度から順に 1, 0.8, 0.2, 0 である．

2.3 善悪推論のシミュレーション

SVO を善行と悪行の生成モデルとして，観察した意思決定結果から相手の SVO を式 1に従ってベイズ推論できるかどうかの計算機シミュレーションを行った．なお，式 2を含めた検討は本稿では行わない．最初の 10 ラウンドでは図 2bに示す鹿狩りゲームをランダムに 2 回ずつ出現させた．11 ラウンドから 15 ラウンドのゲームは図 2bの囚人ジレンマに固定した．性格特性推論の対象となるプレイヤエージェントの真の SVO を 45, 5, -5, -45 度に設定し，自身の SVO に従って 1 から 10 ラウンドは合理的意思決定を，11 ラウンドは非協力（D），それ以降はしっぺ返し戦略による意思決定を行った．観測者エージェントは各ラウンドでランダムに意思決定を行い，$p(SVO) = 0.25$ をプレイ前の事前確率として用い，各ラウンドでのベイズ推論の結果得られた事後確率を次ラウンドの事前確率として用いるベイズ更新を行った．シミュレーション結果を図 2dに示す．グラフより，5 ラウンド程度で真の SVO を高い確率で推論できることがわかる．なお，プレイヤエージェントは 10 ラウンドで裏切り，それ以降しっぺ返しと SVO に関係ない戦略を取っているために，11 ラウンド以降で確率が低下し，観察した行動から性格特性を推論できないことを示している．



(a) 社会的価値志向性（SVO）．横軸が自分の報酬への関心 w_{self}，縦軸が相手の報酬への関心 w_{other} の度合いを示す．本研究では 45 度から-45 度にかけて善（cooperation）から悪（competition）に段階的に変化する SVO を対象にした．

SVO angle and decision

SVO angle	w_{self}	w_{other}	SS'	SH'	HS'	HH'	Row1	Row2	Rc
45	0.71	0.71	4.24	1.41	1.41	1.41	2.83	1.41	S
5	1.00	0.09	3.25	0.17	1.99	1.08	1.71	1.54	S
-5	1.00	-0.09	2.73	-0.17	1.99	0.91	1.28	1.45	H
-45	0.71	-0.71	0.00	-1.41	1.41	0.00	-0.71	0.71	H

(c) 鹿狩りゲーム 3 の利得表に対して計算した，4 種類の SVO のエージェントの期待報酬と合理的意思決定（R_c）．例えば 45 度の SVO では S (Row1) を選択した場合の期待報酬 2.83 が H (Row2) を選択した場合の期待報酬 1.41 を上回るために S を選択する．

(e) 2 × 2 マトリックスゲームを行った UI 画面．

	S'	H'
S	2 , 2	0 , 1
H	1 , 0	1 , 1

鹿狩りゲーム 1

	S'	H'
S	3 , 3	0 , 1
H	1 , 0	1 , 1

鹿狩りゲーム 2

	S'	H'
S	3 , 3	0 , 2
H	2 , 0	1 , 1

鹿狩りゲーム 3

	S'	H'
S	3 , 3	0 , 2
H	2 , 0	2 , 2

鹿狩りゲーム 4

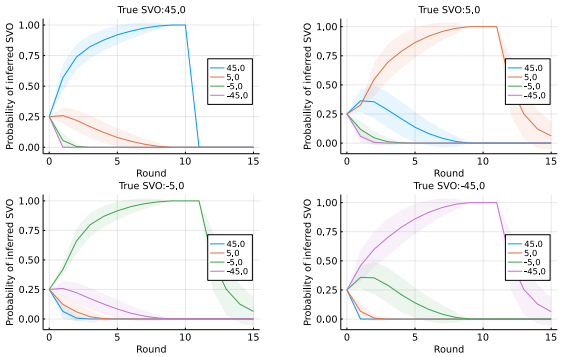
	S'	H'
S	3 , 3	1 , 2
H	2 , 1	2 , 2

鹿狩りゲーム 5

	C'	D'
C	2 , 2	0 , 3
D	3 , 0	1 , 1

囚人のジレンマゲーム

(b) 鹿狩りゲームと囚人のジレンマゲーム．



(d) 4 種類の真の SVO に対して，ベイズ更新により相手の性格（SVO）を推論したシミュレーション．グラフは 10,000 回の平均値．帯は標準誤差，縦軸が相手の SVO の確率，横軸がラウンド．

通常の状態　good 表情　bad 表情

(f) Live2D を使用したエージェントの表情．表情の妥当性は別途検証した．

図 2: 方法

2.4 手順，測定，分析

実験参加者は，図 2eの Web インターフェースを用いて，SVO が 45 度，5 度，-5 度，-45 度のいずれかのエージェントと 15 ラウンドのゲームをプレイするように求められた．エージェントの意思決定はシミュレーションで説明したものと同一であった．S と H の選択に関して位置情報による選択予測ができないように，S と H の位置はランダムに配置した．また，実験参加者に，最初の 10 ラウンドからエージェントの行動意図を見極め，11 ラウンド以降のゲームで高い報酬となる選択をするよう求めた．

なお，式 2の推論に用いられる，各ラウンドの後に表出するエージェントの表情 $p(EE|O, SVO)$ は，式 3で計算される報酬が高い場合に good 表情，低い場合に bad 表情とした（図 2f）．詳細は [8] を参照．

実験参加者の意思決定結果（選択）について，鹿狩りゲームにおける S，囚人のジレンマゲームにおける C を協力として記録した．新規状況において相手の性格特性を考慮した合理的な意思決定ができるかどうかについて，11 ラウンドから 15 ラウンドの協力率につ

いて繰り返し有りの分散分析を行った．

3. 実験結果

15 ラウンドの協力率の平均の推移グラフを図 3に示す．11 ラウンドから 15 ラウンドの協力率について繰り返し有りの分散分析を行った結果，SVO と表情の交互作用は有意傾向（$F(3,364) = 2.315$, $p = .076$,$\eta_p^2 = .019$）であり，SVO 主効果が観測された（$F(3,364) = 17.101$, $p = .000$,$\eta_p^2 = .124$）．また，表情の主効果は確認されなかった（$F(3,364) = 0.001$, $p = .978$,$\eta_p^2 = .000$）．多重比較の結果，SVO 角度 45 度の場合の協力率（$M = 0.567$, $SD = .049$）は-45 度（$M = 0.158$, $SD = .051$）と-5 度（$M = 0.179$, $SD = .054$）よりも高く（-45 度:$p < .05$，-5 度:$p < .05$），SVO 角度 5 度（$M = 0.385$, $SD = .046$）の場合の協力率は-45 度と-5 度よりも高いことが分かった（-45 度:$p < .05$，-5 度:$p < .05$）．また，表情の効果は SVO 角度が 45 度のときのみ観測され，表情表出がある場合（$M = 0.567$, $SD = .049$）の方がない場合（$M = 0.410$, $SD = .054$）よりも協力率が高いことが確認された（$p < .05$）．

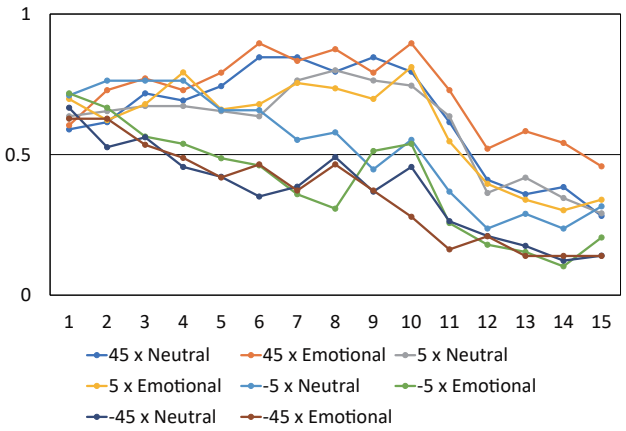


図 3: 実験結果，15 ラウンドの協力率の推移．

4. 議論

SVO の角度に応じた協力率の全順序関係は確認されなかったが，SVO が 45 度，5 度の場合は SVO が-5 度，-45 度の場合よりも高い協力率であった．エージェントの性格特性（SVO）の違いに応じて囚人のジレンマゲームにおける協力率が異なったことは，実験参加者が 10 ラウンドの鹿狩りゲームの間に相手の SVO を見極めて囚人のジレンマゲームにおいて適切に協力と裏切りを選択できたことを意味する．鹿狩りゲームでは S と H の位置をラウンドごとにランダムに変化させていたために，位置情報によってエージェントの選択を予測したのではないことがわかる．また，表情表出があった場合の囚人のジレンマゲームの協力率が表情表出なしの場合よりも高かったことは，式 1の推論に加えて式 2の推論が行われることによって，相手の協力的な性格特性の確信度がより高まり，10 ラウンドの裏切りに対する信用が回復したものと考えられる．

本研究はジレンマを含む社会的状況において，人が相手の善悪を，自他の相対利益を評価をする性格特性のモデルである SVO を生成モデルとして用いたベイズ推論によって推論し，相手が善行を行うか悪行を行うかを予測し，合理的に振舞うことを示した初めての研究である．人が他者の行動の善悪の予測に評判を使うことは知られているが [3]，本研究は反復インタラクションにおいて，相手の評判を形成する認知計算の解明に貢献したと言える．理論的に予測される認知計算のシミュレーション結果は人がベイズ推論を行っていることの傍証になるが，実際に推論を行っているかどうかは確定できないため，今後の研究で検証する．

文献

[1] Ernst Fehr and Urs Fischbacher. The nature of human altruism. *Nature*, Vol. 425, No. 6960, pp. 785–791, oct 2003.

[2] Sandy Bogaert, Christophe Boone, and Carolyn Declerck. Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, Vol. 47, No. 3, pp. 453–480, sep 2008.

[3] Richard D. Alexander. *The Biology of Moral Systems*. Aldine De Gruyter, 1987.

[4] Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, Vol. 1, No. 4, p. 0064, mar 2017.

[5] Yang Wu, Chris L. Baker, Joshua B. Tenenbaum, and Laura E. Schulz. Rational inference of beliefs and desires from emotional expressions. *Cognitive Science*, oct 2017.

[6] György Gergely and Gergely Csibra. Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, Vol. 7, No. 7, pp. 287–292, Jul 2003.

[7] 稲葉美里. 連結による協力問題解決メカニズムの解明. PhD thesis, 北海道大学, 2016.

[8] 長谷川智大, 寺田和憲, セルスドゥメル, フランシスコサントス. 他者モデルの推定による行動の予測と一貫性のある協力行動の実現. 人工知能学会全国大会 (第 36 回) 論文集, 2022.