1L5-3

利己的な相互協調：強化学習主体による不確実な囚人のジレンマ

Mutual cooperation in selfish learners: Prisoner's dilemma under uncertainty

鳥居 拓馬*1　日高 昇平*2
Takuma Torii　Shohei Hidaka

*1東京大学大学院 工学系研究科
School of Engineering, The University of Tokyo

*2北陸先端科学技術大学院大学 知識科学研究科
School of Knowledge Science, Japan Advanced Institute of Science and Technology

A fundamental problem in group decision-making is a conflict between individual and collective rationality: selfish individuals who attempt to maximize his or her own payoff yield sub-optimal outcomes for the group of rational individuals. The present study analyzes the iterated prisoner's dilemma (IPD), known as a model of the social dilemma between individual and collective rationality, and asks what situation can relax it. In view of past psychological and economic literature, we introduce a sort of bounded rationality, in which each player of the IPD does not know the game's payoff structure and therefore each needs to learn it from past experience. Our analysis shows that the two players can establish the mutually cooperative relationship, although they learn to selfishly maximize their own individual payoffs. This result can be interpreted as a relaxation of the social dilemma: bounded individual rationality can also be collective rationality.

1. はじめに

個人と集団の意思決定におけるジレンマ構造は社会性動物の利他的行動の発生，国際政治など多様な文脈で見られるため，囚人のジレンマを基礎とした各種のゲームを扱うことで集団的に最適な意思決定を実現する数理的メカニズムが分析されてきた．囚人のジレンマの拡張として，意思決定が一回性のものではなく繰り返される状況や，複数の意思決定の方策が存在する状況などを考え，集団最適である相互協調の発生条件が論じられている [Axelrod 84, Nowak 92, Nowak 93]．

多くの経験的な研究から，人の意思決定は必ずしもゲーム理論の想定する合理性を満たさないことが指摘されている [Camerer 03, Kahneman 79]．また，囚人のジレンマの心理実験では人間は初回から最適解をとらず，ゲームをプレイするなかで次第に最適解を学習することが知られている [Kahn 93]．これらの知見を踏まえると，合理的な意思決定により生じる個人と集団の間のジレンマが，有限の情報と認知資源に基づく限定合理的な意思決定では生じない可能性がある．

学習するプレイヤ間のゲームの研究は比較的歴史が浅く，ゲームのダイナミクスに及ぼす学習の影響は解明すべき点が多く残されている [Sato 02]．古典的な枠組みでは，各プレイヤが相手の行動やゲームの報酬などすべての情報をえられると仮定した信念学習が研究されてきた．この仮定の下では，仮想プレイ [Brown 51] や相手の行為の模倣（Tit-for-Tat: TFT）[Axelrod 84] などに代表される，相手の行動履歴にもとづく戦略が研究されてきた．他方で，自身のとりうる行動とそれに対する報酬のみからゲームの構造を間接的に推論することを要求する枠組みとして強化学習が研究されている [Sato 02, 鳥居 14, Hidaka in press]（他方で [Sandholm 96] のように相手の情報を参照できると仮定した強化学習もある）．本研究では，限定合理的な意思決定の一種として，それぞれのプレイヤが自身への利得が相手のプレイヤの選択に依存することを知らない状況を考える．この状況では，各プレイヤの選択に対して複数の利得があり，一種の不確実性の下での意思決定の問題として捉えられる．強化学習はこうした不確実な状況の下で漸近的により大きな利得をもたらす（期待利得を最大化する）ことが知られている．本研究では，不確実性下の強化学習によって個人と集団のジレンマが解消されるかを議論する．

連絡先: 鳥居 拓馬，東京都 文京区 本郷 7-3-1，tak.torii@sys.t.u-tokyo.ac.jp

2. 定式化

2.1 繰り返し囚人のジレンマ（IPD）

繰り返し囚人のジレンマ（IPD）では，各プレイヤはふたつの選択肢，協調（Cooperate: C）もしくは裏切り（Defect: D）のうち，いずれかの行動を決定する．ゲームのある時点 t のプレイヤ $i \in \{1,2\}$ の行動を

$$x_{t,i} = \begin{cases} \mathrm{C} & \text{if Cooperate} \\ \mathrm{D} & \text{if Defect} \end{cases} \tag{1}$$

と記す．ふたりのプレイヤの行動（行動組み）は

$$x_t = (x_{t,1},\ x_{t,2}) \in \{\mathrm{C},\mathrm{D}\} \times \{\mathrm{C},\mathrm{D}\}$$

のいずれかとなる．ある時点 t のプレイヤ i と j（$i \neq j$）の行動に対応した報酬の組みは表 1 に記されている．このゲームが繰り返し囚人のジレンマであるには，

$$r(\mathrm{DC}) > r(\mathrm{CC}) > r(\mathrm{DD}) > r(\mathrm{CD})$$
$$\text{and} \quad 2\,r(\mathrm{CC}) > r(\mathrm{CD}) + r(\mathrm{DC})$$

を満たす必要がある．本稿では，$r(\mathrm{CC}) = 3$，$r(\mathrm{CD}) = 0$，$r(\mathrm{DC}) = 5$，$r(\mathrm{DD}) = 1$ を分析した．

2.2 戦略

繰り返し囚人のジレンマの各時点では，両プレイヤの行動の組み合わせは $\mathcal{M} = \{\mathrm{CC},\mathrm{CD},\mathrm{DC},\mathrm{DD}\}$ のいずれかとなる．

表 1: 囚人のジレンマの利得行列

		Cooperate ($x_{t,2}=\mathrm{C}$)	Defect ($x_{t,2}=\mathrm{D}$)
C	($x_{t,1}=\mathrm{C}$)	$r(\mathrm{CC})$	$r(\mathrm{CD})$
D	($x_{t,1}=\mathrm{D}$)	$r(\mathrm{DC})$	$r(\mathrm{DD})$

各プレイヤは，過去の行動組み系列の関数として，次の行動の選択確率を決める．この選択確率を与える関数を戦略と呼ぶ．

本研究では大別して，相手の行動に応じて次の行動を選択する戦略と，相手の行動によらずに次の行動を選択する戦略を分析する．前者の戦略は 1 回前のゲームの帰結に基づき意思決定するため，1 次戦略と呼ぶ．後者の戦略は過去の自らの行動に対する報酬に応じて行動選択をおこなう強化学習の枠組み [Sutton 98] に従うため，強化学習戦略と呼ぶ．本研究では 1 次戦略と強化学習戦略を併せて調べることで，1 次戦略に関する先行研究の知見を強化学習戦略を理解するために応用し，強化学習戦略のもつ限定合理性の性質を分析する．

2.3 1 次戦略

ある結果 y が実現したあとで x が生起する条件付確率を $P(x|y)$ と記すとき，1 次戦略は時点 t のゲームの帰結 $x_t \in \mathcal{M}$ に対する応答として次に協調 C をとる確率 $P(x_{t+1,i}=\mathrm{C}|x_t)$ により表現できる．1 次戦略は，時点 t によらず戦略が一定であるため，これを $P_i(\mathrm{C}|y)$ と書く．

本研究では，先行研究 [Axelrod 84, Nowak 92, Nowak 93] で分析された 7 つの 1 次戦略と比較する．経験的な分析から，他のすべての 1 次戦略に優る万能な戦略は見つかっていないが，しっぺ返し戦略 TFT（Tit-for-Tat）は相対的に高い平均利得をえることが知られている．一部の戦略を除いて，1 次戦略は基本的に直前の相手の行動に反応して次の行動を選択する．

簡潔に各 1 次戦略について述べる．ALLC はかならず C をとるという戦略である．ALLD はかならず D をとるという戦略である．RAND は常に C と D を等確率（1/2）でとる戦略である．GRIM は，履歴がすべて CC の場合を除いて D をとる戦略である．TFT（Tit-for-Tat）は最初は C をとり，以後は相手の直前の行動を模倣するという戦略である．GTFT（Generous TFT）は TFT とほぼ同じだが，相手が D のときでも一定の確率（1/3）で C をとるという戦略である．WSLS（Win-Stay, Lose-Shift）は「負け」なら行動を切り替え，「勝ち」なら前と同じ行動をとるという戦略である．

2.4 強化学習戦略

囚人のジレンマでは各プレイヤの利得が相手の選択に依存しており，合理的主体や 1 次戦略の一部は各プレイヤがその利得行列を既知であることを前提に意思決定をおこなう．一方，本研究では限定合理性の一種として，各プレイヤがゲームの利得行列（表 1）の情報をもたずに行動を選択する場合を考える．この場合，プレイヤ 1 からは 2 つの行動の組み合わせ $(x_{t,1}, x_{t,2}) = (\mathrm{C}, \mathrm{D})$ と (C, C) を区別できず，自分の選んだ行動 C に対して未知なる相手の行動（D または C）に依存した利得 $r(\mathrm{CD})$ または $r(\mathrm{CC})$ をうけとる．したがって，過去の自分の選択とそれに対する自分への利得だけの関数として次の時点における選択確率を与える戦略を考える．

強化学習戦略では，ある時点 t において，過去 K 時点前までの両プレイヤの行動の履歴

$$X_t = (x_{t-1},\ x_{t-2}, \ldots,\ x_{t-K})$$

に対して，時間に応じて割引した累積報酬

$$R_{i,x}(X_t) = \sum_{k=1}^{K} \alpha_i^k\, \delta(x, x_{t-k,i})\, r_{t-k,i} \tag{2}$$

を考える．式 (2) において，

$$\delta(x,y) := \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$$

であるため，$R_{i,x}(X_t)$ はプレイヤ i が時点 $t-k$ で行動 $x_{t-k,i}$ を選択した場合に，その行動に対する報酬 $r_{t-k,i}$ を時間に依存する割引係数 α_i^k で重み付けたものの総和を表す．累積報酬は記憶保持率パラメータ $0 \leq \alpha_i \leq 1$ の関数であり，α_i が大きいほど，過去の利得をより高く評価する．強化学習戦略は累積報酬 $R_{i,x}(X_t)$ と鋭敏性パラメータ $\beta_i \geq 0$ をもつロジスティック関数として以下のように与えられる：

$$P_i(x|X_t) = \frac{\exp[\beta_i\, R_{i,x}(X_t)]}{\sum_y \exp[\beta_i\, R_{i,y}(X_t)]}. \tag{3}$$

鋭敏性 β_i が大きいほど累積報酬の高い行動をより高い確率で選択し，$\beta_i = 0$ の場合は C か D を確率 1/2 で選択する．

3. 分析方法

繰り返し囚人のジレンマの各時点は 4 つの状態 CC，CD，DC，DD をもち，その状態の系列に対して次の状態への遷移確率が一意に定まる．このような確率過程はマルコフ過程と呼ばれ，とくに K 時点前までの状態のみに依存する確率過程は K 次マルコフ過程と呼ばれる．強化学習戦略は式 (3) が与える K 次マルコフ過程として分析できる．

任意の t に対し X_t のとりうる状態系列の集合を X とする．履歴長 K の強化学習戦略は，条件付確率 $P(X_{t+1}|X_t)$ を与え，初期状態によらず[*1]，十分な時間経過ののち，$|X| = 4^K$ の状態系列上の定常分布へ収束する．定常分布は遷移確率 $P(X_{t+1}|X_t)$ を用いて，$P(X_{t+1}) = P(X_t)$ の解として計算できる．

履歴長 K の強化学習戦略に対し，4^K の状態系列上の定常分布 $Q(X) = Q(x_1, x_2, \ldots, x_K)$ の和をとり 1 状態の確率変数とする（周辺化する）ことで，$\mathcal{M} = \{\mathrm{CC}, \mathrm{CD}, \mathrm{DC}, \mathrm{DD}\}$ 上の周辺定常分布を定義する．本研究の目的すなわち個人と集団の意思決定における最適性を論じるために，相互協調の実現確率 $Q(\mathrm{CC})$ および相互裏切りの実現確率 $Q(\mathrm{DD})$ を分析する．

4. 強化学習戦略による相互協調

IPD の枠組みにおいて，両プレイヤが個別に個人の利得を最大化した結果として，集団全体としても利得を最大化することが可能だろうか．前述のとおり，強化学習戦略は基本的に個人的な期待利得を最大化するよう過去の履歴に基づき行動を選択する（個人の利得の最大化）．IPD の利得行列（表 1）によると，ふたりの平均利得（集団の利得）を最大化するのは相互協調 CC である．そこで，まず強化学習戦略同士の IPD において相互協調の発生確率が高くなる条件および個人と集団の意思決定の最適性が一致する場合があることを確認する．とくに強化学習戦略において重要な役割を担うパラメータのひとつ，記憶保持率 α_1，α_2 の役割について分析をおこなった．

*1 Perron-Frobenius の定理より，非周期的かつ既約な場合，定常分布が初期状態に依存しない

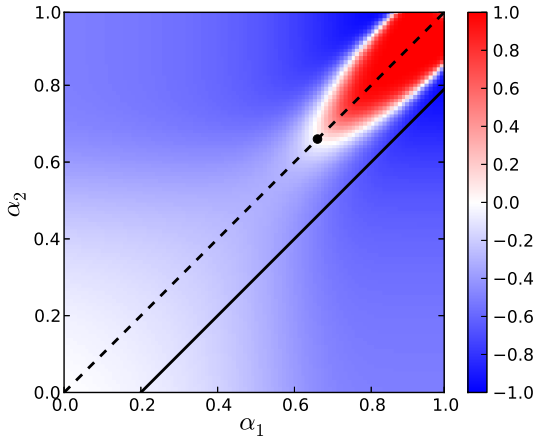

図 1: 相互協調確率と相互裏切り確率の差

図 1 は $\alpha_i = 0.0, 0.01, \ldots, 0.98, 0.99$ の範囲で相互協調 CC と 相互裏切り DD の発生確率の差 $Q(\mathrm{CC}) - Q(\mathrm{DD})$ を示している．もし集団の最適性がより頻繁に実現されていれば $Q(\mathrm{CC}) > Q(\mathrm{DD})$ となる．この図から，DD に比べて CC が高い確率で実現されるには，厳密に等しい記憶保持率でなくてもよいが，双方が同程度に高い記憶保持率をもつことが重要であると考えられる．言い換えれば，一方だけが低い記憶保持率をもつ場合，相互裏切りの確率が高くなる．

5. 1 次戦略との比較分析

前節の分析では，学習のパラメータである記憶保持率が同程度に高いとき，強化学習戦略同士の囚人のジレンマにおいて相互協調 CC が発生することを確認した．しかし，強化学習戦略がどのように振る舞うことで相互協調を達成しているかは明らかではない．強化学習戦略の意思決定は複雑であるため，本節では各種の 1 次戦略の定常周辺分布と比較することで，強化学習戦略による相互協調の発生メカニズムを分析する．

5.1 定常周辺分布の分析

強化学習戦略と 1 次戦略を比較するために，定常周辺分布を分析する．定常周辺分布は $Q(\mathrm{CC})+Q(\mathrm{CD})+Q(\mathrm{DC})+Q(\mathrm{DD}) = 1$ を満たすため，4 次元単体上に表現できる．また，今回の分析ではプレイヤの交換に対する対称性を考慮して $Q(\mathrm{CD})$ と $Q(\mathrm{DC})$ を区別しない．この場合，定常周辺分布は 3 次元単体にあり，平面 $(x, y) = (Q(\mathrm{CC}) - Q(\mathrm{DD}), Q(\mathrm{CD}) + Q(\mathrm{DC}))$ 上の点として一意に表現できる．図 2 では，7 つの 1 次戦略すべてのペアに対して定常周辺分布を算出し，この単体面上に白抜きの点（○）で示した．1 次戦略の複数の組み合わせが同じ点に重なるため，白抜きの各点の側に 1 次戦略のペアを表示している．図 2 に示す周辺分布の単体面は 3 つの頂点からなり，それぞれ (ALLC, ALLC)，(ALLD, ALLD)，(ALLC, ALLD) である．単体面の中央 $(0, 1/2)$ に RAND がある．

5.2 強化学習戦略の 1 次戦略による近似

強化学習戦略の結果は縁無しの点（●）で示されている．1 次戦略とは異なり，図 2 の単体面上の $Q(\mathrm{CD}) + Q(\mathrm{DC}) \leq 1/2$ に分布している．強化学習戦略の各点の色は両プレイヤの記憶保持率の差の絶対値 $|\alpha_1 - \alpha_2|$ から設定されており，青色の点では差がほぼ 0 で，緑色，赤色，橙色と変わるにつれて差が大きいことを表す．強化学習戦略が $\alpha_1 = \alpha_2 = 0.8$ あたりから (ALLC, ALLC) に漸近することを除いて，1 次戦略の分布

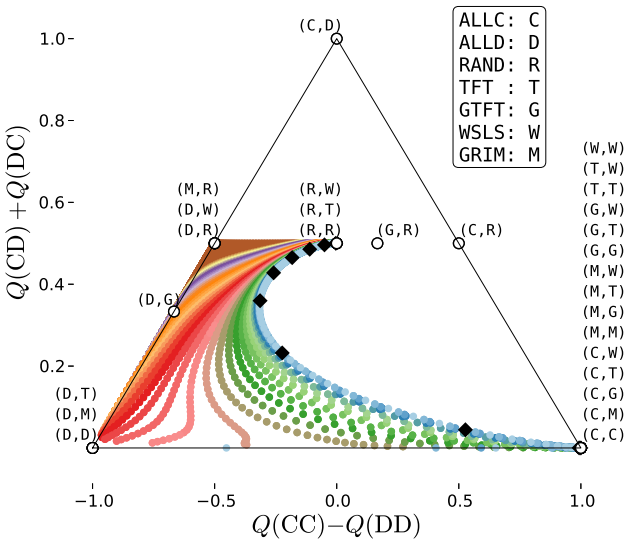

図 2: 定常周辺分布の単体面

から強化学習戦略の行動選択のメカニズムに関する洞察をえることは難しい．そこで，先行研究の 7 つに限らず，1 次戦略クラスに含まれる戦略の中で，強化学習戦略の周辺確率分布をもっともよく近似する 1 次戦略を逆算し，この 1 次戦略の性質を分析した．この分析では，初期に確率 $P_0(\mathrm{C})$ で協調し，遷移確率 $(P(\mathrm{C}|\mathrm{CC}), P(\mathrm{C}|\mathrm{CD}), P(\mathrm{C}|\mathrm{DC}), P(\mathrm{C}|\mathrm{DD}))$ にしたがう同一の 1 次戦略同士の周辺確率が目標とする強化学習戦略の周辺確率と最小二乗誤差をもつよう遷移確率を定めた*2．

図 2 の黒い菱形の点（◆）は，プレイヤ間で記憶保持率が等しい場合 $\alpha_1 = \alpha_2 = 0.1, 0.2, \ldots, 0.9$ の強化学習戦略同士の周辺分布（図 1 の破線）をもっともよく近似する 1 次戦略を示している．また，図 3 に近似 1 次戦略の遷移確率を $\alpha := \alpha_1 = \alpha_2$ の関数として示した．図 3 の各点は遷移確率の平均値，実線は $\alpha \pm 0.04$ の窓による移動平均を表す．図 3 では $P(\mathrm{C}|\mathrm{CC}), P(\mathrm{C}|\mathrm{DD})$ のみを図示した．

図 3 から，もっともよく近似する 1 次戦略が $\alpha = 0.65$ 付近を境に切り替わっていることがわかる．記憶保持率 $\alpha < 0.65$ の範囲では $P(\mathrm{C}|\mathrm{CC})$ が低く，$P(\mathrm{C}|\mathrm{DD})$ が高い．一方で，$\alpha \geq 0.65$ の範囲では $P(\mathrm{C}|\mathrm{CC})$ が高く，$P(\mathrm{C}|\mathrm{DD})$ が低い．

図 3 の結果を，確率 $P(\mathrm{C}|\mathrm{CC})$ と $P(\mathrm{C}|\mathrm{DD})$ のパタンから，記憶保持率 α の区間に対応する 4 つのフェーズ I，II，III，IV に便宜的に分類して説明する．フェーズ I は $0.0 \leq \alpha < 0.3$ の区間に対応し，RAND（ランダム選択）と類似した戦略がみられる．フェーズ II は $0.3 \leq \alpha < 0.6$ の区間に対応し，前回の状態が CC でも DD でも裏切る確率のほうが高いという点で確率的なゆらぎのある ALLD（いつでも裏切り）に近い戦略と解釈できる．

フェーズ III は $0.6 < \alpha < 0.65$ の区間に対応し，強化学習戦略は合理的プレイヤのように相互協調 CC から裏切りに転じるが，一方で相互裏切り DD の場合には協調に転じる．先行研究の代表的な戦略にはこのような 1 次戦略は含まれていないが，ALLD（相互協調時にも裏切る）と WSLS（相互裏切りに対して協調する）を混合した戦略とみなせる．この戦略

*2 強化学習戦略の周辺分布を近似する 1 次戦略は一般には一意に定まらず，無数にありえる．ここでは 1 次戦略のパラメータ空間上の一様乱数で 1000 セット の初期値を与え，それらの初期値に対してえられた最小二乗解のうち，二乗誤差が 10^{-13} 以下の解の平均パラメータを分析した．

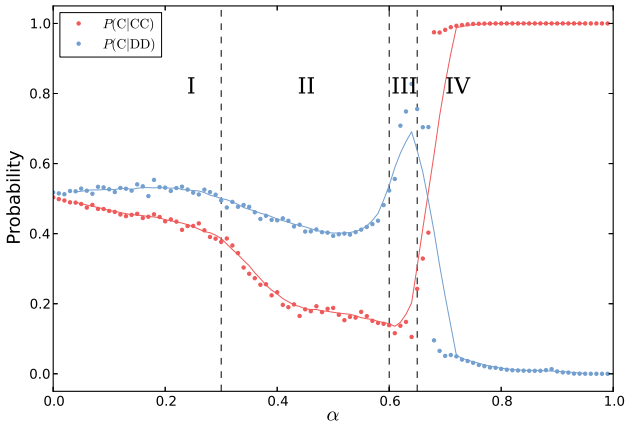

図 3: 強化学習戦略（$\alpha_1 = \alpha_2$）を近似する 1 次戦略

は，結果的に状況を相互協調を裏切りへ，相互裏切りを協調へと撹乱する戦略と解釈できる．

最後に，フェーズ IV は $\alpha > 0.65$ の区間に対応し，強化学習戦略は TFT（しっぺ返し）戦略に類似した振る舞いを示すと解釈できる．換言すると，相互協調時にはそれを継続し，相手が裏切る場合には自分も裏切り返す．実際には強化学習戦略は直接的に相手の行動を参照して行動を選択していないが，累積報酬を通じて間接的に TFT と類似の行動パタンを示すことが可能になったと考えられる．また $\alpha = 0.65$ 付近の前後で，こうした近似 1 次戦略の分岐が起こり，この分岐点は相互協調確率と相互裏切り確率の差が 0 になる点（図 1 の破線上の黒丸の点 $\alpha_1 = \alpha_2 = 0.65$）と一致する．

以上の分析から，対称な強化学習戦略において相互裏切りよりも相互協調が高い確率で発生した背景には，記憶保持率の低くゆらぎのある裏切り戦略から，1 次戦略でいうしっぺ返し戦略（TFT）への振る舞いの定性的な変化があると考えられる．

また，$\alpha_1 = \alpha_2 + 0.2$ を満たす場合を，異なる記憶保持率をもつ強化学習戦略の典型とみなし 1 次戦略で近似して分析した．その結果，記憶保持率が小さくほぼランダムな戦略から，記憶保持率が十分に大きく過去の履歴に依存した選択をおこなう戦略への変化を除いて，定性的な変化は見られなかった．

6. 考察

古典ゲーム理論 [Neumann 44] や進化的手法 [Axelrod 84, Nowak 92, Nowak 93] では，十分な情報に基づく強い推論によって，個人の合理性を超えた相互協調は発生せず（古典ゲーム理論），多様な戦略への適応という進化的な視点を導入することで相互協調の発生を模索してきた（進化的分析）．一方，本研究では，互いに相手の行動に関する情報を与えられず，しかし個人の利得を最大化する強化学習戦略を考え，自己の期待利得を最大化することがそのまま集団としての期待利得を最大にする場合があることを示した．具体的には，実効的に同程度の長さの行動履歴に基づき行動を選択する強化学習戦略の間では相互協調が高い確率で発生することを示した．また，1 次戦略による近似から，この相互協調の発生は学習によるしっぺ返し戦略のような戦略の発生に伴うこともわかった．利得が確率的で不確実であることが，強化学習同士の対戦において，個人と集団のジレンマの解消，つまり個人と集団の利得の最大化が一致した主要因であると考えられる．

一般に囚人のジレンマのように行動と帰結のすべてが事前に明らかであるような状況は少なく，むしろ特定の問題に関する利害構造はその問題に取り組んだ結果から推測しなければならないことが多いと考えられる．この観点からは，ゲーム理論的な合理性よりも，むしろ自分の過去にとった行動とそれに対する利得のみから，統計的により期待利得の高い行動を選ぶ強化学習戦略は自然であると考えられる．本分析の含意は，多くの現実の状況でみられる情報の限定性や不確実性により，結果として，個人と集団の意思決定が共通の望ましい選択を実現している可能性を示唆している．

参考文献

[Axelrod 84] Axelrod, R.: *The Evolution of Cooperation*, Basic Books, New York (1984)

[Brown 51] Brown, G.: Iterative solution of games by fictitious play, in Koopmans, T. ed., *Activity Analysis of Production and Allocation*, pp. 374–376, John Wiley & Sons, New York (1951)

[Camerer 03] Camerer, C. F.: Behavioural studies of strategic thinking in games, *Trends in Cognitive Sciences*, Vol. 7, No. 5, pp. 225–231 (2003)

[Hidaka in press] Hidaka, S., Torii, T., and Masumi, A.: Which types of learning make a simple game complex? (in press)

[Kahn 93] Kahn, L. M. and Murnighan, J. K.: Conjecture, uncertainty, and cooperation in prisoner's dilemma games, *Journal of Economic Behavior and Organization*, Vol. 22, pp. 91–117 (1993)

[Kahneman 79] Kahneman, D. and Tversky, A.: Prospect theory: An analysis of decisions under risk, *Econometrica*, Vol. 47, No. 2, pp. 263–291 (1979)

[Neumann 44] Neumann, von J. and Morgenstern, O.: *Theory of Games and Economic Behavior*, Princeton University Press (1944)

[Nowak 92] Nowak, M. and Sigmund, K.: Tit-for-tat in heterogeneous populations, *Nature*, Vol. 355, pp. 250–253 (1992)

[Nowak 93] Nowak, M. and Sigmund, K.: A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game, *Nature*, Vol. 364, pp. 56–58 (1993)

[Sandholm 96] Sandholm, T. and Crites, R.: Multiagent reinforcement learning in the iterated prisoner's dilemma, *Biosystems*, Vol. 37, No. 1–2, pp. 147–166 (1996)

[Sato 02] Sato, Y., Akiyama, E., and Farmer, J.: Chaos in learning a simple two-person game, *Proceedings of the National Academy of Sciences*, Vol. 99, No. 7, pp. 4748–4751 (2002)

[Sutton 98] Sutton, R. and Barto, A.: *Reinforcement Learning: An Introduction*, MIT Press (1998)

[鳥居 14] 鳥居 拓馬, 日高 昇平, 真隅 暁: 学習あり繰り返し囚人のジレンマにおける協調行動の発生, 第 28 回人工知能学会全国大会予稿集 (2014), 4H1-3