社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ICS－147 (13)
2007／3／15

二重ジレンマ状態での協調行動の誘発に関する意思決定手法の検討

和田 志保美* 鈴木 恵二†

* 公立はこだて未来大学大学院システム情報科学研究科
† 公立はこだて未来大学システム情報科学部

概 要

囚人のジレンマでは、パレート効率的な結果を導き出す協調行動ではなく、ナッシュ均衡をもたらす裏切り行動が確認されることが良く知られている。本論文では、囚人のジレンマをプレーするかしないかという問題自体が囚人のジレンマとしての要素を持つ「二重囚人のジレンマ・ゲーム」を用いてヒトと強化学習エージェントの実験結果を比較する。実験結果から、最終的に協調を選択する場合はヒトも強化学習エージェントもどのようなゲームをプレーするかに関する選好を示さなかった。一方、最終的に裏切りを選択する場合は、どのようなゲームをプレーするかに関する選好を強化学習エージェントが示さなかったのに対し、ヒトでは通常の囚人のジレンマをプレーすることを好んだ。このことから、適切な設計を行うことによってマシンエージェントが協調を選択し、パレート効率的な結果に到達しやすくなることが示唆される。

キーワード： 二重囚人のジレンマ, 囚人のジレンマ, 強化学習, 選好

An examination of emerging cooparation under double-bind dilemma

Shihomi WADA* **Keiji SUZUKI†**

* Graduate School of Future University-Hakodate, System Information Science
†Future University-Hakodate, Departmemt of System Information Science

Abstract

In this paper, we compare the experimental results of human agent with that of programmed agent using double-bind prisoner's dilemma game, which ordinary prisoner's dilemma game is nested into another dilemma, that is, a player has to decide firstly if s/he will play prisoner's dilemma or not. We find that when cooperation is chosen in the second stage, both human and programmed agents show any preference in what kind of 2 × 2 game they play. On the other hand, human agents are willing to play ordinary prisoner's dilemma game when they choose defection in the second stage, though programmed agents show any preference in the first stage. This suggests that programmed agents may be cooperators and get close to the Pareto optimal equilibrium in an appropriate setting.

Keywords : double-bind prisoner's dilemma game, prisoner's dilemma game, reinforcement learning, preference

*g3105006@fun.ac.jp
†suzukj@fun.ac.jp

生産者はジレンマに直面している。より多く生産すれば収入が増えるものの、価格破壊を起こすことがある。2006 年 3 月に過剰生産された 892 トンの牛乳が廃棄処分となった (Fukaya [4])。大量生産をするかしないか、生産者にとってこれはジレンマである。

公共財投資ゲームを用いて Hauert, Monte, and Hofbauer [5] は社会的ジレンマが生じない簡単なメカニズムを示した。彼らは協調と裏切りの他に年金生活という第三の選択肢を用いた。年金生活者は公共財投資ゲームに参加しないものの、全員が裏切り合った場合よりも多い利得で全員が協調し合う場合よりも少ない利得を年金として受け取る。このことで協調 – 裏切り – 年金生活がジャンケンのような関係になり、社会的ジレンマを解消する。

しかし年金生活者となることを選択するのは現実的ではない。さらに Orbell and Dawes [6] が示すようにプレーするかしないかを選ぶ権利がある。Hauert et al. [5] は年金生活者が公共財投資ゲームに参加しないと設定しているが、「プレーをしない」ことと「ゲームに参加しない」ことは同義ではない。良くも悪くも誰もが他者とかかわりを持たずには生活できない。個々人はプレーをするかしないかを選ぶ権利があるのと同様、各自の人生を全うする権利がある。そこで我々はプレーをするかしないかを選ぶことの出来る簡単なメカニズムを提案する。

先程の牛乳の過剰生産を例に取ってみよう。年金生活者になることは現実的ではない。過剰生産か適度生産かの 2 選択ではなく、現実味のある第三の選択肢は設定できないか。ここで例えば加工生産を考える。過剰生産者は裏切り者であり、適度生産者は協力者である。加工生産者は、加工に要する機材などの初期投資が多い分だけ弱い存在であるが、生産量に関するジレンマからは逃れることができる。通常の囚人のジレンマ (PD) に、支配される戦略である加工生産者を付加した新しい 3 × 3 は、依然 PD である。

本論文では、二重囚人のジレンマ (DbPD) での被験者実験の結果とシミュレーション実験の結果を比較する。DbPD は通常の PD が PD の中に入れ子になったものである。つまり、DbPD は PD をプレーするかしないかが PD となったゲームである。

次章でゲームのモデルを説明し、被験者実験とシミュレーション実験の詳細を述べる。その後、実験の仮説を説明し、最後に結果と考察をまとめる。

1 実験

この章では最初に実験に用いたゲームの構造について述べる。その後、被験者実験の手順と、シミュレーション実験の詳細について述べる。

1.1 ゲームの構造

本研究では通常の囚人のジレンマ・ゲーム (PD) と、Wada and Suzuki [8] で紹介されている二重囚人のジレンマ・ゲーム (DbPD) を用いた。ここでは最初に DbPD の利得表を示し、次にどのように DbPD がプレーされるかを説明する。

利得表 表 1 に PD の利得表を、表 2 に DbPD の利得表を載せる。

表 1: 囚人のジレンマ (PD) の利得表

	D	C
D	2 2	1 5
C	5 1	4 4

表 2: 二重囚人のジレンマ (DbPD) の利得表

	D	C	E
D	2 2	1 5	1 5
C	5 1	4 4	1 5
E	5 1	5 1	3 3

表 1 と表 2 に示された PD と DbPD の利得表を比較すると、DbPD の利得表は PD の利得表に選択肢 E を付加しただけであることがわかる。つまり DbPD の利得表の左上部分が PD の利得表であることに注意されたい。選択肢 E は通常であれば支配される戦略として抹消される選択肢である。よって PD の利得

表に選択肢 E を付加することは、ナッシュ均衡およびパレート効率に対していかなる影響も与えない。このことから DbPD は PD の一種であることがわかる。

次に、DbPD のルールを説明する。

二重囚人のジレンマ・ゲームのルール　1 回の DbPD において、プレイヤーは 2 回の意思決定を行う。

第一段階では、各プレイヤーは第二段階で自分が選択しない選択肢 α を 1 つ決める。各プレイヤーの選択肢 α は同時に公開され、第二段階に移る。第二段階では、相手が出さない選択肢を考慮した上で、各プレイヤーは最終的に選択する選択肢 β を 1 つ決める。どのプレイヤーも第二段階では第一段階とは異なる選択をしなくてはならない ($\beta \neq \alpha$)。各プレイヤーの第二段階での選択肢 β が同時に公開され、これによって各プレイヤーは表 1 に示されたとおりの利得を獲得する。

1 ラウンドあたりの DbPD の流れを、表 3 にまとめておく。

表 3: 1 ラウンドあたりの DbPD の流れ

	ラウンド開始直前
1	プレイヤーは自分の対戦相手を確認する。
	第一段階
2	各プレイヤーは第二段階で使わない選択肢 α (α = C, D, E) を決める。
3	プレイヤーは相手がどの選択肢を第二段階で使わないかを同時に知る。
	第二段階
4	各プレイヤーは第二段階で使う選択肢 β (β = C, D, E ただし $\beta \neq \alpha$) を決める。
5	プレイヤーは最終的に相手の選んだ選択肢を同時に知る。
6	表 2 に基づいてプレイヤーは利得を受け取る。

選択肢が 2 つの場合、被験者がその選択肢を選びたくて選んだのか、あるいは被験者が他方の選択肢を選びたくなかったから選んだのか、実験者は判別することができない。このことは、PD であれば、被験者が協調したくなかったから裏切ったのか、あるいは裏切りたくて裏切ったのかの区別がつかないことを意味する。DbPD では選択肢 α は選びたくないから選ばなかった選択肢であり、選択肢 β は選びたくて選んだ選択肢であることがわかる。

第一段階の意思決定で被験者が選択肢 D を第二段階では使わないと決めた場合、この被験者は PD をプレーしたくないことを意味する。また、支配戦略である選択肢 D を用いる機会を第二段階に残しておくことは、第二段階で PD をプレーする可能性が高まり、結果として自分の利得を低くする危険性が伴う。この観点に立つと、被験者は別の問題に直面していることがわかる。ジレンマに陥るか、陥らないか、それがジレンマである。

二重囚人のジレンマという名前の所以はこの点にある。ゲームの構造については以上である。次に被験者実験とシミュレーション実験についての詳細を述べる。

1.2　被験者実験の手続き

本節では被験者実験の手続きについて説明する。

被験者実験は 2006 年 3 月 23 日に長野県松本市にある尾の上湯旅館にて、湧源クラブの旅行のオプションとして行われた。数理科学に興味のある 36 名の被験者が自発的に参加した。

被験者実験は第一セッションと第二セッションに分かれている。第一セッションでは PD を 4 ラウンド行い、第二セッションでは DbPD を 5 ラウンド実施した。どちらのセッションでも、各ラウンドは 1 ゲームから成り立つ。

実験を開始する前に、被験者は二重の円形に座る。内側の円と外側の円のそれぞれの被験者が向かい合い、二人組を作り、この組内でゲームが行われる。ラウンドが終了すると、一方円に座っている被験者が実験者の指示に従って横に移動し、対戦相手を擬似ランダムに従って変更する。このように繰り返しのないゲームが行われた。

被験者の意思表示は、じゃんけんの手を使って全員同時に行われる。第一セッションでは、紙は協調を、石は裏切りとして用いた。第二セッションでは、石は協調を、はさみは被支配選択を、紙は裏切りとして用いた。第二セッションの第一段階では、被験者は全員同時に両手を出すことで意思表示をするよう指示された。もし被験者が石とはさみを出した場合、この被験者は第二段階で裏切りは選択できない。第二段階では、第一段階で出した手のどちらか一方を同時に

引っ込めることで意思表示を行う。最終的に出ている手が、その被験者の第二段階の選択肢となり、表2に示した利得表に定められた利得を得てラウンドは終了となる。

1.3 シミュレーション実験の手続き

ここではシミュレーション実験に用いたエージェントの説明をした後、実験の詳細について述べたい。

エージェント シミュレーション実験では7種類のエージェントを用いた。うち6種類は常に決まった戦略を用いる固定エージェントであり、残る1種類は強化学習に基づいたエージェントである。固定エージェント (α, β) は選択肢 α を第一段階の選択肢とし、選択肢 β を第二段階の選択肢とする。

強化学習エージェントのモデルは Feltvich [3] を参考にした。第 t ラウンドの第一段階で、強化学習エージェントは情報 Φ に応じて選択肢 α (α = C, D, E) を第二段階で使用しないことに対する非負の傾向 $q_1^t(\alpha|\Phi)$ を有する。Feltvich [3] に従って第 t ラウンドの第一段階における状況 Φ での全ての選択行動に対する傾向の総和 $Q_1^t(\Phi)$ は、選択肢 α (α = C, D, E) を第二段階で使用しないことに対する非負の傾向の総和とした。すなわち $Q_1^t(\Phi) = q_1^t(C|\Phi) + q_1^t(D|\Phi) + q_1^t(E|\Phi)$ である。

第2ラウンド以降は、前のラウンドでの獲得利得 π_{t-1} を単純に加算することで選択肢 α (α = C, D, E) を第二段階で使用しないことに対する非負の傾向 $q_1^t(\alpha|\Phi)$ を更新する。

$$q_1^{t+1}(\alpha|\Phi) = \begin{cases} (1-f)\cdot q_1^t(\alpha|\Phi) + \pi_t \\ \qquad (\alpha\text{が選択されたとき}) \\ (1-f)\cdot q_1^t(\alpha|\Phi) \\ \quad (\alpha\text{が選択されなかったとき}) \end{cases}$$

ここで f は忘却パラメータであり、シミュレーション実験では $f = 0.1$ とした。

情報 Φ において第 t ラウンドの第一段階で選択肢 α を選択する確率は、$p_1^t(\alpha|\Phi) = q_1^t(\alpha|\Phi) \,/\, Q_1^t$ である。

第 t ラウンドの第二段階において、強化学習エージェントは情報 Φ の元で選択肢 β (β = C, D, E ただし $\beta \neq \alpha$) を選択する傾向 $q_2^t(\beta|\Phi)$ を有する。ここで $q_2^t(\alpha|\Phi) = 0$ とする。第 t ラウンドの第二段階での選択に対する傾向の総和 $Q_2^t(\Phi)$ は $Q_2^t(\Phi) = q_2^t(C|\Phi) + q_2^t(D|\Phi) + q_2^t(E|\Phi)$ となる。

第2ラウンド以降は、前のラウンドでの獲得利得 π_{t-1} を単純に加算することで選択肢 β (α = C, D, E) を選択することに対する非負の傾向 $q_2^t(\beta|\Phi)$ を更新する。

$$q_2^{t+1}(\beta|\Phi) = \begin{cases} (1-f)\cdot q_2^t(\beta|\Phi) + \pi_t \\ \qquad (\beta\text{が選択されたとき}) \\ (1-f)\cdot q_2^t(\beta|\Phi) \\ \quad (\beta\text{が選択されなかったとき}) \end{cases}$$

第一段階と同様に f を忘却パラメータとし、実験では $f = 0.1$ とした。また、情報 Φ において第 t ラウンドの第二段階で選択肢 β を選択する確率は、$p_2^t(\beta|\Phi) = q_2^t(\beta|\Phi) \,/\, Q_2^t$ である。

なお、平均利得が3であることから、$q_1^0(\alpha \in S|\Phi) = 3$ とし、$q_2^0(\beta \in S|\Phi) = 3$ とした。付録に強化学習エージェントのパラメータをまとめておく。

常に決まった選択を行う固定戦略エージェントが、強化学習エージェントの学習が完了するまで繰り返し対戦するのを1ブロックとする。6体の固定戦略エージェントがそれぞれ1,000体の強化学習エージェントと対戦することにより、実験は全部で6,000ブロック行った。

実験は 3.20 GHz の Intel Celeron CPU と 224 MB RAM を搭載した eMachines J3042 を用いて実施した。

1.4 仮説

まず、合理的で利己的なプレイヤーが DbPD をどのようにプレーするかを確認したい。合理的で利己的なプレイヤーは、第二段階で起き得ることを踏まえた上で第一段階の選択を決定する。支配される戦略 E が付加されただけで、DbPD は依然 PD である。よって合理的で利己的なプレイヤーは選択肢 D を第二段階で選択する (β = D) と考えられる。以上より、次の2つの仮説が立てられる。

仮説 1 選択肢 D は第二段階で選択される選択肢 β である。

仮説 2 第一段階で選択肢 D が選択されることはない。

これら 2 つの仮説から、$p_2^t(C) + p_2^t(E) = 0$ であり、かつ $p_1^t(C) + p_1^t(E) = 1$ であると推測できる。合理的で利己的なプレイヤーにとって重要なことは、第二段階で選択肢 D を選択することである。このようなプレイヤーにとっては、選択肢 C と選択肢 E はどちらも選択肢 D によって支配される選択に過ぎないため、無差別である。このことから、以下の仮説を立てることができる。

仮説 3 第一段階において、選択肢 C と選択肢 E はどちらも α となる可能性があり、両者には差がない。

これら 3 つの仮説はいずれもゲーム理論で言うところの合理的で利己的なプレイヤーを想定している。これとは別に一方で、直感的に次のような合理的で利己的なプレイヤーを想定することも可能である。第二段階で通常の PD をプレーすれば、裏切り合うことになり、期待利得を低めることになる。PD をプレーしないことで、期待利得を高めることができる。このような直感的に合理的で利己的なプレイヤーを仮定することにより、次のような仮説が立てられる。

仮説 4 プレイヤーは第二段階で PD をプレーすることを回避する。

ここで $p_1^t(E) \neq 1$ であることは、$\alpha \neq$E であることを意味している。仮説 2 と 4 を合わせて考えると、$\alpha =$C であると予想できる。しかし、このことはいささか奇妙である。なぜなら $\alpha =$C であると、第二段階においてプレイヤーは本来のパレート効率的な結果にたどり着くことができないのである。

次章で実験結果を確認することで、これらの仮説を吟味する。

2 結果と総括

ここでは先に実験結果を示し、後半に総括を述べたい。

2.1 被験者実験の結果

被験者実験における PD の結果を表 4 に、DbPD の結果を表 5 に示す。

表 4: PD での選択比率の推移

ラウンド	C	D
1	66.67 %	33.33 %
2	50.00 %	50.00 %
3	50.00 %	50.00 %
4	52.78 %	47.22 %

表 5: DbPD での選択の分布 (被験者)

第一段階 (α)	第二段階 (β)			合計
	D	C	E	
D	——	34	3	37
C	27	——	3	30
E	63	49	——	112
合計	90	83	6	179

被験者のうち PD を知っていた被験者が 3 名混じっており、これらの被験者が PD をプレーすることを退屈に感じたため、PD は 4 ラウンドで打ち切った。それでも PD において協調行動が確認された割合は 66.67% から 52.78% に落ち込んだ。

DbPD は 36 名の被験者で 5 ラウンド行われたが、第 2 ラウンドで 1 名のデータが未記入であったため、データの総数が 179 となっている。第二段階で裏切りを選択した被験者と協調を選択した被験者を比較する。最終的に裏切った被験者のうち 7 割が第一段階で E を選択しているのに対し、最終的に協調を選んだ被験者が第一段階で E を選択した割合は 6 割弱にとどまっている。このことから、協調的な被験者は支配戦略と被支配戦略のどちらを放棄するかに対する選好が弱いことがわかる。一方で、第二段階で裏切りを選択する被験者は、PD をプレーすることを好む傾向にある。

よって以下の結論が導き出せる。

結果 1. 最終的に裏切る被験者は PD をプレーしたがる。

結果 2. 協調的な被験者はどの 2 × 2 ゲームを第二段階で行うかに対する選好が弱い。

実験終了後に被験者に口頭で確認をしたのだが、どの被験者もDbPDの利得表の中にPDが埋め込まれていることには気が付かなかった。さらに、PDをプレーするかしないかがジレンマとなることに気が付いた被験者も皆無だった。

2.2 シミュレーション実験の結果

強化学習エージェントが固定戦略エージェントとの対戦でどのような戦略を獲得したのかを表6にまとめる。

表6から次の2つの結果が導き出せる。

> **結果3.** 強化学習エージェントの学習結果は、固定戦略エージェントの第二段階の選択によって決まる。

> **結果4.** 強化学習エージェントは第二段階でどの2 × 2をプレーするかに関する選好を持たない。

結果2と結果4より、仮定3は正しいと言えるが、このことは結果1に反する。

2.3 総括

本論文では、通常のPDをプレーするかしないかがPDとなっていることでPDが二重に入れ子状態になったDbPDを用いて、被験者実験の結果とシミュレーション実験の結果を比較した。

これらの比較により、注目すべき結果として、類似点と相違点が見つかった。第二段階で協調行動を選択する場合、被験者もプログラムエージェントもどの2 × 2ゲームをプレーするかに対する選好を示さなかった。一方で、第二段階で裏切る場合も実際にプレーする2 × 2ゲームに対する選好をプログラムエージェントは持たないのに対し、第二段階で裏切る被験者はPDをプレーすることを好んだ。

ここでマイノリティゲームについて考えたい。マイノリティゲームでは、適切な条件設定のされたエージェントが過去の利得に関する履歴に基づいた意思決定をすることにより、全体として適切な選択行動が実現する (Challet and Zhang [1])。マイノリティゲームの原型では2選択課題であったが、Chow and Chau [2] は3つ以上の選択肢を用いたマイノリティゲームを実施し、その場合でも全体として適切な選択行動が実現することを確認した。

再びDbPDに戻って考えたい。今回のエージェントシミュレーションでは、獲得した利得に基づくフィードバックによってのみパラメータが更新される。ここで第一段階でどのような選択をした場合にどんな利得が得られたかに関する履歴を用いると、DbPDがマイノリティゲームの要素を持つと考えられる。このような履歴情報を実験で与えることは可能である。もちろんDbPDをプレーするプレイヤー達の目的は第二段階終了時に得られる利得であり、どれだけの人数が特定の2 × 2ゲームに参加したかではない。しかし被験者実験の結果とシミュレーション実験の結果の比較をする以上、このような追加実験を行うことは有意義であると考えられる。

表 6: シミュレーション実験の結果

条件 (C, D)

α	β D	C	E	合計
D	——	19	31	50
C	466	——	19	485
E	443	22	——	465
合計	909	41	50	1,000

条件 (C, E)

α	β D	C	E	合計
D	——	253	21	274
C	245	——	29	274
E	234	218	——	452
合計	479	471	50	1,000

条件 (D, C)

α	β D	C	E	合計
D	——	179	0	179
C	328	——	1	329
E	353	139	——	492
合計	681	318	1	1,000

条件 (D, E)

α	β D	C	E	合計
D	——	242	29	271
C	273	——	29	302
E	200	227	——	427
合計	473	469	58	1,000

条件 (E, C)

α	β D	C	E	合計
D	——	176	0	176
C	353	——	0	353
E	349	122	——	471
合計	702	298	0	1,000

条件 (E, D)

α	β D	C	E	合計
D	——	26	37	63
C	448	——	23	471
E	434	32	——	466
合計	882	58	60	1,000

参考文献

[1] Challet, D., and Zhang, Y. C., "Emergence of Cooperation and Organization in An Evolutionary Game", *Physica A*, **246**, pp. 407 - 418 (1997).

[2] Chow, F. K., and Chau, H. F., "Multiple Choice Minority Game", *Physica A*, **319**, pp. 601 - 615 (2003).

[3] Feltvich, N., "Reinforcement-based vs. Belief-based Learning Models in Experimental Asymmetric-information Games", *Econometrica*, **68 (3)**, pp. 605 - 641 (2000).

[4] Fukaya, T., "Milk's popularity on the wane", *Daily Yomiuri*, April 18 (2006).

[5] Hauert, C., Monte, S. D., and Hofbauer, J., "Volunteering as Red Queen Mechanism for Cooperation in Public Goods Games", *Science*, **296**, pp. 1129 - 1132 (2002).

[6] Orbell, J. H., and Dawes, R. M., "Social welfare, cooperators' advantage, and the option of not playing the game", *American Sociological Review*, **58**, pp. 787 - 800, (1993).

[7] Wada, S., and Suzuki, K., "Double-bind Prisoner's Dilemma Game", *Conference Proceedings of North American Computational Social and Organization Sciences (NAACSOS) 2006 CD-ROM.*

[8] Wada, S., and Suzuki, K., "How to Reach Pareto Optimum in Double-bind Prisoner's Dilemma Game", *Proceedings of Joint 3rd International Conference on Soft Computing and Intelligent Systems and 7th International Symposium on Advanced Intelligent Systems (SCIS & ISIS 2006) CD-ROM.*

付録 強化学習エージェントのパラメータ

全体を通じて用いられるパラメータ	
t	ラウンド
Φ	情報
f	忘却パラメータ (当論文では $f = 0.1$ とする)
π_t	第 t ラウンドで獲得した利得
第一段階で用いられるパラメータ	
α	エージェントが第二段階では使わないと決定した選択肢 (α = C, D, E)
$q_1^t(\alpha\|\Phi)$	第 t ラウンドにおける状況 Φ で選択肢 α を第二段階で使用しない傾向
$Q_1^t(\Phi)$	第 t ラウンドの第一段階における状況 Φ での全ての選択行動に対する傾向の総和 ($Q_1^t(\Phi) = \sum_{\alpha\in\Phi}(q_1^t(\alpha\|\Phi))$
$p_1^t(\alpha)$	第 t ラウンドの第一段階で選択肢 α を選ぶ確率 ($p_1^t = q_1^t(\alpha\|\Phi)/Q_1^t(\Phi)$)
第二段階で用いられるパラメータ	
β	エージェントが最終的に用いる選択肢 (β = C, D, E ただし $\beta \neq \alpha$)
$q_2^t(\beta\|\Phi)$	第 t ラウンドにおける状況 Φ で選択肢 β を第二段階で使用する傾向
$Q_2^t(\Phi)$	第 t ラウンドの第二段階における状況 Φ での全ての選択行動に対する傾向の総和 ($Q_2^t(\Phi) = \sum_{\beta\in\Phi}(q_2^t(\beta\|\Phi))$
p_2^t	第 t ラウンドの第二段階で選択肢 β を選ぶ確率 ($p_2^t = q_2^t(\beta\|\Phi)/Q_2^t(\Phi)$)