強化学習モデルは人間行動をどの程度説明するか？
――均衡が1度だけ移動する経済実験を例に――†

小　川　一　仁[1)]

Ⅰ　はじめに

1970年代後半から進展した実験経済学や行動経済学の研究成果の多くは，主流派の経済学やゲーム理論の予想と大きく異なるものであった。最後通牒ゲーム実験（Güth, et al.［1982］など）や独裁者ゲーム実験（Kahneman, et al.［1986］）での理論的分配結果との大きな乖離，リスクに対する態度分析（Kahneman and Tversky［1979］, Tversky and Kahneman［1981］［1986］など）における期待効用理論の破綻とプロスペクト理論による修正，公共財供給実験（Ledyard［1995］など）における協力の達成とその崩壊などがこれらの好例である[2)]。以上の結果は経済主体の完全合理性仮定に関する否定的根拠となった。

次に現れたのは，従来の理論が想定する水準の合理性を持たない経済主体を想定した理論を再構築するという動きだった。たとえば，フェール・シュミットに代表される不平等回避型効用関数（Fehr and Schmidt［1999］），カメレール型（Camerer［2003］）およびロス型の強化学習モデル（Roth and Erev［1995］, Erev and Roth［2001］による意思決定モデル，マッケルヴィーらによる質的応答均衡モデル（McKelvey and Palfrey［1995］），スタールらによるレベルkの意思決定理論（Stahl［1998］, Stahl and Haruvy in press）などである。これらのモデルは従来の経済理論の結果では説明できなかった人間行動をうまく説明できることが多い。

上述した意思決定モデルは，静的な環境において行われた経済実験の結果を従来の理論よりも適切に説明できる。実験の間中ステージゲームの均衡が変化しない状況では新たなモデルは人間行動をより適切に説明する。この状況では，実験の間中，効用関数ないし利潤関数が変化しない。

しかし，実験の中でステージゲームの均衡が変化する状況で，これらのモデルがどの程度人間行動を説明するのかが，検討されたことはほとんどない。均衡が移動する状況は，現実世界において日常的に発生している。たとえば，市場均衡は日々刻々と変化している。このような現実の状態を考慮すると，均衡が移動するケースにおいて，これらの新たに登場したモデルが人間行動を十分説明するかどうかを検討することには意義があると思われる。

本稿では，実験の途中で均衡が1度だけ変化する場合に，どのような意思決定モデルが説明力が高いのかを検討する。実験の前半と後半で均衡が正反対の位置に存在する経済実験で，エレブ―ロス型の強化学習モデルとそのバリエーションが，ゲーム理論の均衡予測と比べて，人間行動をどの

---

† 筆者は学部生の頃から問題関心がひとつに定まらない経済学徒であった。このような筆者を八木紀一郎先生は常に温かく見守ってくれた。先生の定年退職に当たって，これまでの学恩に対して心から感謝の意を表したい。
1）本稿は科学研究費補助金（基盤研究(C)課題番号：19530281 代表：二村英夫）による研究成果の一部である。
2）一方で，ダブルオークション実験における効率性の高さは市場理論の予想を支持している（Smith［1962］）。

程度説明するかを検討する。

具体的には，公共財供給実験（Ledyard［1995］など）とそのバリエーションである地域通貨導入型の公共財供給実験を取り上げる。この実験は，地域通貨を事前に使用したことがあるという体験が，その後の地域社会での個々人の協力姿勢にどのような影響を与えるかを検討するために実施された。この実験が考案・実施された背景には1990年代の後半から2000年代初頭において，日本においても地域通貨の設立，運用が盛んになった事実がある（西部［2002］，室田［2004］，三浦［2008］）。ちなみに日本国内で現在までに発行された地域通貨は約652種類[3]であるという。

後述するように，この実験の均衡は前半と後半で異なる。均衡の違いに対して被験者がどのように反応したのか。また反応の仕方はどのようなルールに基づくと考えられるか。本稿ではこれらの点について，計算機上に被験者実験と同じ環境を構築し，その中における仮想的プレイヤの振る舞いを手がかりに，被験者がどのように行動していたと解釈できるかを検討する。

Ⅱ 被験者実験の概要[4]

今回実施した被験者実験では，通常の通貨を用いる公共財供給ゲームと地域通貨を用いる公共財供給ゲームの2つを用いた。各ゲームは2ステージから構成される。1ステージ目では被験者は独立同時にゲームに参加するかどうかを選択する。不参加を選択した者は，利得40を受け取り，そのラウンドの意思決定は終了する。参加を選択した者は2ステージ目において投資を行い，以下で述べる利得を獲得する。

被験者iのラウンドtでの利得について述べる。通常の通貨を用いる公共財供給ゲームでは$profit_{i,t}=40-k_i+\frac{R}{m}\sum_{i=1}^{m}k_i$である。ただし，$R=1.8$である。$k_i$は被験者$i$の投資量，$m$はゲームの参加人数で，2または3である（$m=1$の時，ゲームは実施されない）。

通常の効果を用いる公共財供給ゲームでは，各プレイヤは2段階目のどの情報集合においても投資しない。つまり$n=0$であり，各プレイヤは40の利得を獲得する。また，第1段階目で検討される公共財供給ゲームへの参加については，参加する場合もしない場合も部分ゲーム完全均衡戦略になりうる（参加しない場合でも利得40を獲得できる）。さらに，（参加する，投資しない）は全てのプレイヤにとって弱支配戦略である。部分ゲーム完全均衡がプレイされる場合，または全てのプレイヤが弱支配戦略を用いる場合，全てのプレイヤの投資額は0であり，利得は40である。しかし，全てのプレイヤが（参加する，全て投資する）を用いると，全プレイヤの利得は$1.8\times40=72$となり，均衡での利得と比べてパレート優位である。部分ゲーム完全均衡はプレイヤのただ乗りを誘発し，非効率な結果となる。

3）http://www.cc-pr.net/list/，2009年3月30日アクセス。休止，廃止したものも含む延べ数。また，重田（2005）によると地域通貨は以下の特徴を持つ。もちろん，全ての特徴を持つ地域通貨もあれば，1部しか持たないものもある。1．特定の地域内（市町村など），あるいはコミュニティ（商店街，町内会，NPO）などの中においてのみ流通する，2．市民ないし市民団体（商店街やNPOなど）により発行される，3．無利子またはマイナス利子である，4．人と人をつなぎ相互交流を深めるリングとしての役割を持つ，5．価値観やある特定の関心事項を共有し，それを伝えていくメディアとしての側面を持つ，6．原則的に法定通貨とは交換できない。

4）被験者実験のより詳しい設定については，二村他（2009）を参照せよ。

一方，地域通貨を用いる公共財供給ゲーム[5] では $profit_{i,t}=d\times(40-k_i)+\frac{R}{m}\sum_{i=1}^{m}k_i$ である。ただし $d=0.5$，$R=1.8$ である。このゲームでは，唯一の部分ゲーム完全均衡が存在する。それは，全てのプレイヤが（参加する，全て投資する）を選択するというものである。これは効用関数を 1 階微分すれば明らかである。

被験者実験は 2 つのトリートメント，LC First トリートメントと PG First トリートメントから構成される。それぞれのトリートメントは前半と後半の 2 つのセッションからなり，各トリートメントの各セッションでは，地域通貨ゲーム 30 ラウンドまたは公共財供給ゲーム 30 ラウンドを行う。また，両方のゲームで被験者の初期保有は 40 とし，投資額は 0，10，20，30，40 のうちからしか選べないこととした。初期保有は毎ラウンド与えられ，次ラウンド以降への持ち越しは禁止された。

LC First トリートメントでは，前半のセッションで，グループを固定して地域通貨ゲームを 30 ラウンド行い，後半のセッションで公共財供給ゲームを地域通貨ゲームのときと同じグループで 30 ラウンド行う。地域通貨ゲームを行う際，後半のセッションの内容を伝えず，このゲームが終わってから公共財供給ゲームの説明を行う。またゲームをプレイする 3 名 1 組のグループは地域通貨ゲーム開始時にランダムに選ぶ。このとき，匿名性を保つために，グループの他のメンバーが誰であるかは伏せられた。

公共財供給ゲームは，地域通貨ゲームと同じグループで行うが，このことも地域通貨ゲーム終了後に伝えられる。謝金は，2 つのセッションで得た合計得点を 0.6 倍し，それに 700 を加えた値に等しい額の日本円を支払った。

PG First トリートメントは前半のセッションで公共財供給ゲームを行い，後半のセッションで地域通貨ゲームを行う点を除き，LC First トリートメントと同じである。

被験者実験は，広島市立大学情報処理センターで実施した。2009 年 2 月 23 日に LC First トリートメントを，2009 年 2 月 24 日に PG First トリートメントを実施した。1 トリートメントあたりの実施時間は約 2 時間で，被験者は同大学国際学部の学部生計 51 人（LC first 24 人，PG first 27 人[6]）であった。全ての被験者は 1 つのトリートメントにのみ参加した。謝金は最低が 2168 円，最高が 3120 円，平均 2680 円であった。この謝金水準は被験者が他所で稼得する場合の機会費用を考えると妥当な水準である。なお，実験アプリケーションは z-Tree を用いた（Fischbacher［2007］）。

5）地域通貨の諸特徴のうち，「利子がマイナス」である点をモデル化したのが地域通貨導入型の公共財供給ゲームである。

6）LC first トリートメントでは男性が 2 人，女性が 22 人，PG first トリートメントでは男性が 5 人，女性が 22 人であった。国際学部の特性上，女性の参加者が多いが，性差は実験結果に影響したとは言えなかった。

Ⅲ　被験者実験の結果

本節では被験者実験の結果を簡単に解説する[7)]。以下では，LC first トリートメントにおける第1，第2セッションをそれぞれ LC1，PG2 と表示し，PG first トリートメントにおける第1，第2セッションをそれぞれ PG1，LC2 と表示することにする。

表1は各セッションで，第1段階で「参加する」を選んだ被験者の第2段階における平均拠出額の前半 15 ラウンドの平均，後半 15 ラウンドの平均，全ラウンドの平均を示している。また，図2はそれぞれセッション LC1 と LC2 およびセッション PG1 と PG2 の各ラウンドにおいて，第1段階で「参加する」を選んだ被験者の第2段階における平均拠出額の推移を示したものである。

これらの結果から，1．LC1 の拠出額は PG2 の拠出額よりも高い，2．PG1 の拠出額は LC2 の拠出額よりも低い，3．LC1 の拠出額は LC2 の拠出額よりも高い，4．PG1 の拠出額は PG2 の拠出額よりも低いことが読み取れる。

これらの直観は計量経済学的分析によって支持される。各個人の性別，学年，その個人が属する

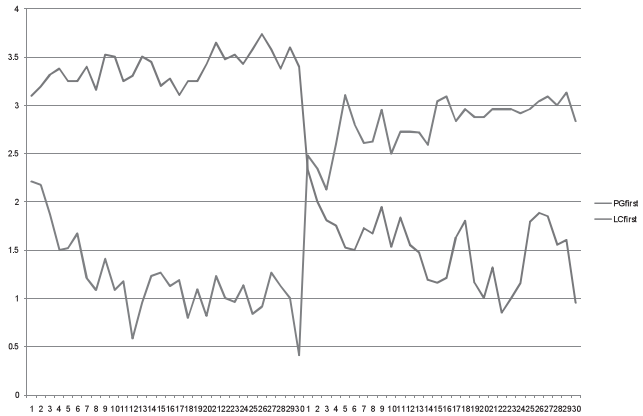

図1　PG first トリートメントと LC first トリートメントにおける拠出量の推移（全 60 ラウンド分）

出典：二村他［2009］

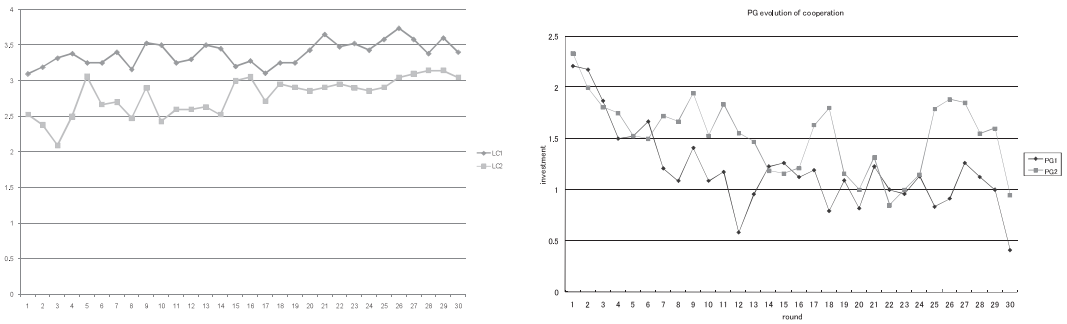

図2　公共財供給ゲームの拠出量の推移（左図：地域通貨使用，右図：通常の通貨使用）

出典：二村他［2009］

7）被験者実験の詳しい結果は二村他［2009］を参照せよ。

表１ 経済実験での投資量の推移（計算機実験とあわせるために10で除した）

Treatment	session	first half	second half	all rounds
LC first	LC1	3.318	3.444	3.381
PG first	LC2	2.601	2.965	2.785
PG first	PG1	1.396	0.992	1.194
LC first	PG2	1.666	1.383	1.524

出典：二村他［2009］

グループの今期のゲーム参加人数を独立変数とし，ランダムエフェクトトービット分析を行ったところ，以下のことがわかった。

ゲームがPGであれば拠出量が有意に低まること，逆に言えばゲームがLGのときに拠出量が有意に高まることが明らかになった。また，LC2セッションの拠出率がLC1セッションのそれに比べて有意に低いことも明らかになった。つまり，最初のセッションでPGをプレイするよりも２つ目のセッションでPGをプレイする方が拠出量は高い。一方，ラウンドが進むごとにLCセッションでは拠出量が高まる，すなわち均衡に近づくことが明らかになった。PGセッションではラウンドを経るごとに拠出量が減少すること，すなわち均衡に近づくことがわかった。これは既存の公共財供給実験の結果と合致している。

Ⅳ．計算機実験における仮想プレイヤの行動原理について

今回のシミュレーションでは２種類のプレイヤを想定した。強化学習プレイヤ１と強化学習プレイヤ２である。

１．強化学習

最初に，全ての種類のプレイヤに共通する意思決定アルゴリズムである，強化学習について述べる。本節でのアルゴリズムはErev and Roth［2001］，Iwasaki, et al.［2007］を参考に構築した。

期待利得を計算するために，起こりうる全ての事象を計算する。まず，３人がゲームに参加し，投資する場合が計125通り存在する。２人が参加する場合には，当該プレイヤが参加するかどうかで２通りに分けられる。当該プレイヤが参加しない場合は25通り，参加する場合は50通り存在する。３人中１人しかゲームに参加しない場合は，公共財供給ゲームが実施されない。これは３通り存在する。最後に全員が参加しない場合は１通りだけ存在する。

強化学習アルゴリズムについて説明する。最初に初期値を与える。各投資量のプロペンシティの初期値は $q_{i,1}(k)=A_{i,1}$ で与えられる。$k\in[0,4]$ は投資量である。i はプレイヤ番号，t（ここでは１）はラウンド番号である。

ただし，PG firstトリートメントのとき，

$$A_{i,1}=\frac{\sum_{l=0}^{4}\sum_{j=0}^{4}\sum_{m=0}^{4}\left[(4-i)+\frac{1.8}{3}\times(i+j+m)\right]+\sum_{l=0}^{4}\sum_{j=0}^{4}\left[(4-i)+\frac{1.8}{2}\times(i+j)\right]\times 2+25\times 4+4\times 3+4}{125+25+50+3+1}。$$

LC first トリートメントのとき，

$$A_{i,1}=\frac{\sum_{i=0}^{4}\sum_{j=0}^{4}\sum_{m=0}^{4}\left[0.5\times(4-i)+\frac{1.8}{3}\times(i+j+m)\right]+2\times\sum_{i=0}^{4}\sum_{j=0}^{4}\left[0.5\times(4-i)+\frac{1.8}{2}\times(i+j)\right]\times2+25\times4+4\times3+4}{125+50+25+3+1}$$

である。

これらの値は各事象が等しい確率で発生するとした場合のプレイヤの期待利得に等しい。利得の散らばりを表す値として，$S_{i,t}$ を毎ラウンド計算する。最初に初期値について計算しておく。PG first トリートメントにおける $S_{i,t}$ の初期値は

$$S_{i,1}=\frac{\sum_{i=0}^{4}\sum_{j=0}^{4}\sum_{m=0}^{4}\left|(4-i)+\frac{1.8}{3}\times(i+j+m)-A_{i,1}\right|+\sum_{i=0}^{4}\sum_{j=0}^{4}\left|(4-i)+\frac{1.8}{2}\times(i+j)-A_{i,1}\right|\times2+25\times|4-A_{i,1}|+|4-A_{i,1}|\times3+|4-A_{i,1}|}{125+25+50+3+1}。$$

一方，LC first トリートメントでは

$$S_{i,1}=\frac{\sum_{i=0}^{4}\sum_{j=0}^{4}\sum_{m=0}^{4}\left|0.5\times(4-i)+\frac{1.8}{3}\times(i+j+m)-A_{i,1}\right|+\sum_{i=0}^{4}\sum_{j=0}^{4}\left|0.5\times(4-i)+\frac{1.8}{2}\times(i+j)-A_{i,1}\right|+25\times|4-A_{i,1}|+|4-A_{i,1}|\times3+|4-A_{i,1}|}{125+25+50+3+1}$$

で計算される。

投資量は以下の確率分布に従って選択される。ここで λ は学習の強さを表す変数で，0 のときは選択確率は全て等しくなり，学習の影響を受けない[8]。λ が大きくなるにつれて最大の q 値を持つ拠出額が選ばれる確率が急激に大きくなる。

$$Prob_{i,t}(k)=\frac{exp(\lambda q_{i,t}(k)/S_{i,t})}{\sum_{j=0}^{4}exp(\lambda q_{i,t}(j)/S_{i,t})}$$

投資量のプロペンシティは以下の式に従って更新される。ただし $W_{i,t}=\frac{1}{Num_{i,t}(k)+\frac{\eta}{125+25+50+3+1}+1}$ である。$Num_{i,t}(k)$ はプレイヤ i がラウンド t までに投資量 k を選択した回数である。η はパラメータであり，この値が大きくなると $W_{i,t}$ が小さくなる。q 値に占める $profit_{i,t}$ の大きさが小さくなることを示している。プロペンシティは以下のように更新される[9]。

$$q_{i,t+1}(k)=\begin{cases}q_{i,t}(k)(1-W_{i,t})+profit_{i,t}\ W_{i,t} & \cdots\text{投資量 k が選ばれたとき}\\ q_{i,t}(k) & \cdots\text{投資量 k が選ばれなかったとき}\end{cases}$$

また，$A_{i,t}$ および $S_{i,t}$ は以下のように更新される。η が大きくなると $\mathrm{profit_{i,\,t}}$ および $\mathrm{S_{i,\,t}}$ の加重が大きくなる。

$$A_{i,t+1}=profit_{i,t}\frac{t+\eta}{t+\eta+1}+\left(1-\frac{t+\eta}{t+\eta+1}\right)A_{i,t}$$

8）q は毎回の選択とその結果得られる利得によって変化するものの，λ が 0 であるため，その変化は拠出量の選択には全く影響しない。すなわち，一様分布である。

9）選択しなかった拠出量のプロペンシティが一定の割合で小さくなる（＝忘却していく）タイプの学習も存在する。これは頻繁に選ばれる拠出量が，将来的により頻繁に選ばれるために導入する手法の 1 つである。

$$S_{i,t+1}=S_{i,t}\frac{t+\eta}{t+\eta+1}+|A_{i,t}-profit_{i,t}|\times\left(1-\frac{t+\eta}{t+\eta+1}\right)$$

次に１ステージ目の意思決定について述べる。ゲームに参加するかどうかの選択は強化学習で選択され，全てのプレイヤに共通の意思決定アルゴリズムである。$participation=1$ のときゲームに参加し，$participation=0$ のときゲームに参加しない。選択確率は以下のように計算される。

$$GameProb\,(participation)=\frac{exp\,(\lambda qp_{i,t}\,(participation)/Spart_{i,t})}{\sum_{j=0}^{1}exp\,(\lambda qp_{i,t}\,(j)/Spart_{i,t})}$$

$Spart_{i,t}$ および $Apart_{i,t}$ の初期値は $S_{i,1}$ および $A_{i,1}$ と等しい。$Spart_{i,t}$ および $Apart_{i,t}$ は以下のように更新される。

$$Apart_{i,t+1}=profit_{i,t}\frac{t+\eta}{t+\eta+1}+\left(1-\frac{t+\eta}{t+\eta+1}\right)Apart_{i,t},$$

$$Spart_{i,t+1}=Spart_{i,t}\frac{t+\eta}{t+\eta+1}+|Apart_{i,t}-profit_{i,t}|\times\left(1-\frac{t+\eta}{t+\eta+1}\right)。$$

選択されたプロペンシティの更新は以下の通りである。

$$qp_{i,t+1}\,(participation)=qp_{i,t}\,(participation)\times(1-Wpart_{i,t})+Wpart_{i,t}\times profit_{i,t}$$

$$Wpart_{i,t}=\frac{1}{Num_t\,(participation)+\dfrac{\eta}{125+25+50+3+1}+1}$$

選択されなかったプロペンシティの更新は行われない。すなわち，$qp_{i,t+1}\,(participation)=qp_{i,t}\,(participation)$ である。

２　戦略選択を行うプレイヤの行動原理

強化学習プレイヤ２は強化学習プレイヤ１に比べて複雑な意思決定アルゴリズムを有する。この種のプレイヤは，強化学習に従って投資量を選択するか，後述するヒル・クライミングに従って投資量を選択するかを毎ラウンド強化学習で選択する。

戦略が選ばれる確率は以下の式で決まる。$l=1$ のとき強化学習，$l=2$ のときヒル・クライミングが選択される。

$$StrategyChoiceProb_{i,t}\,(l)=\frac{exp\,(\lambda\times SCq_{i,t}\,(l)/SCS_{i,t})}{\sum_{j=1}^{2}exp\,(\lambda\times SCq_{i,t}\,(j)/SCS_{i,t})}$$

戦略選択のプロペンシティは以下のように更新される。

$$SCq_{i,t+1}\,(l)=\begin{cases}SCq_{i,t}\,(l)(1-SCW_{i,t})+profit_{i,t}\times SCW_{i,t} & \cdots\text{戦略 } l \text{ が選ばれたとき}\\ SCq_{i,t}\,(l) & \cdots\text{戦略 } l \text{ が選ばれなかったとき}\end{cases}$$

ここで SCW は以下のように計算される。

$$SCW_{i,t}=\frac{1}{Num_t\,(l)+\dfrac{\eta}{125+25+50+3+1}+1}$$

$Num_t\,(l)$ はラウンド t までに戦略 l が選択された回数である。

３　ヒル・クライミング

ヒル・クライミングは強化学習プレイヤ２のみが持つ意思決定アルゴリズムである。ヒル・クラ

イミングが選択される回数が5回目までのとき，投資量0から4がランダムに選択される。6回目以降は $max[WV_{i,t}(k)]$ なる $k \in [0, 4]$ が選択される。WV の更新は以下の通りである。

$$WV_{i,t+1}(k)=\begin{cases}\left(1-\dfrac{1}{(t^{\omega}+1)}\right)WV_{i,t}(k)+\dfrac{1}{(t^{\omega}+1)}profit_{i,t}\cdots\text{投資量 } k \text{ が選ばれたとき}\\ WV_{i,t}(k)\cdots\text{投資量 } k \text{ が選ばれなかったとき}\end{cases}$$

ただし，$WV_{i,1}(k)=A_{i,1}$ である。t はラウンドである。ω は *profit* のウェイトを決める変数である。この値が大きくなると *profit* のウェイトが小さくなり，毎ラウンドの結果が反映されにくくなる。

以上をまとめよう。最初に強化学習プレイヤ1について述べる。強化学習プレイヤ1は最初にゲームに参加するかどうかを強化学習アルゴリズムに従って決定する。参加が決定された場合各ラウンドの最初に各投資量の選択確率を計算する。次にくじを引き，その結果に従って投資量を選ぶ。ゲームの結果としての利得が与えられると，それを用いて投資量の選択確率を計算するのに用いるプロペンシティおよび参加の意思決定をするのに用いるプロペンシティを更新する。

次に強化学習プレイヤ2について述べる。強化学習プレイヤ2も最初にゲームに参加するかどうかを強化学習アルゴリズムに従って決定する。かれは強化学習に基づく意思決定アルゴリズムとヒル・クライミングに基づく意思決定アルゴリズムの2つを持ち，どちらのアルゴリズムが選ばれるかは強化学習で選択される。選択された意思決定アルゴリズムに従って投資量を選択肢，ゲームの結果としての利得が与えられる。それを用いて選択された意思決定アルゴリズムのプロペンシティを更新する。以上をまとめると図3および図4のようになる。

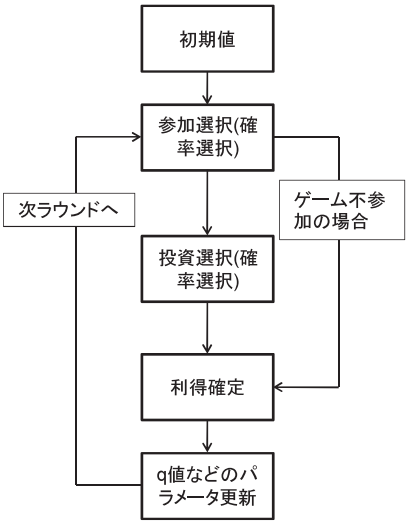

図3　強化学習1のフローチャート
出典：筆者作成

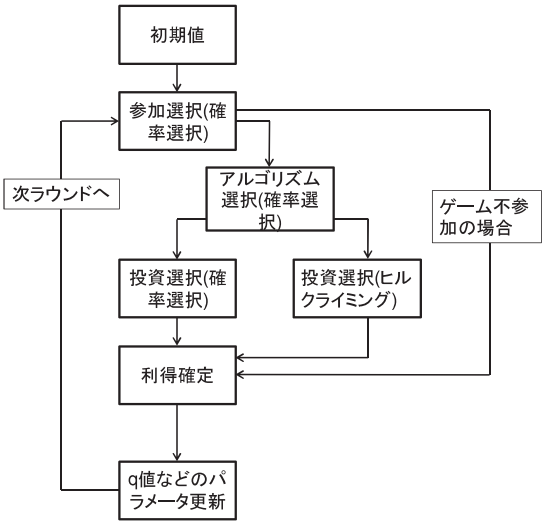

図4　強化学習2のフローチャート
出典：筆者作成

V　計算機実験の概要

計算機実験はLC firstトリートメントとPG firstトリートメントの両方に対して行った。今回，仮想プレイヤごとに学習アルゴリズムを割り当てた。そのため，（3人とも強化学習プレイヤ1），

（２人が強化学習プレイヤ１，１人が強化学習プレイヤ２），（１人が強化学習プレイヤ１，２人が強化学習プレイヤ２），（３人が強化学習プレイヤ２）の４通りが存在する。これらをそれぞれ，3REL，2REL，1REL，0REL と呼ぶ。

これら１通りに対して $\lambda \in [0, 7.8]$，$\eta \in [190, 220]$，$\omega \in [0, 3]$ の範囲で各パラメータの組み合わせに対して，乱数のシードを変えて 100 回計算機実験を行った[10]。ただし，（３人とも強化学習プレイヤ１）のケースでは，ω は不要なパラメータである。このケースでは λ と η の各パラメータの組み合わせに対して 100 回計算機実験を行った。

被験者の意思決定をどの程度説明しているかを測定するために，MSD（Mean Squared Deviation）を導入する（Selten［1998］）。$MSD(\lambda, \eta, \omega)_t=(OC_t-SC_t)^2$ で計算され，被験者行動と完全に一致する場合，最小値の 0 をとる。また，最大値は 16 である[11]。ここで OC_t はラウンド t において被験者実験で観察された平均拠出量である。また，SC_t はラウンド t での拠出量の計算機実験 100 回分の平均値である。最後に各パラメータにおける MSD は $MSD(\lambda, \eta, \omega)=\sum_{t=1}^{60} MSD(\lambda, \eta, \omega)_t$ で得られる。この値を用いて被験者の行動を分析する。

1　計算機実験の結果と分析

最初に，全プレイヤがゲームに参加した場合のナッシュ均衡を選択し続けた場合の MSD と全プレイヤがランダムに意思決定したときの MSD を示す。これらは計算機実験で得られた MSD に対する比較対象となる。

PG first トリートメントにおいて毎ラウンド全プレイヤがナッシュ均衡を選択し続けた場合，最初の 30 ラウンドではゲームに参加し，投資しないので，0 を選択する。後半の 30 ラウンドではゲームに参加し，全て投資するので 4 を選択する。この場合の MSD は 118.411 となる。

PG first トリートメントにおいて全プレイヤがゲームへの参加も含めてランダムに意思決定をした場合は（３人とも強化学習プレイヤ１）で $\lambda=0$ と同じである。このときの MSD は 96 から 97 であった。

一方，LC first トリートメントでの均衡拠出量は PG first トリートメントとは逆になる。最初の 30 ラウンドではゲームに参加し，全て投資するので 4 を選択する。後半の 30 ラウンドではゲームに参加し，投資しないので，0 を選択する。この場合の MSD は 84.824 となる。

PG first トリートメント同様，LC first トリートメントにおいて全プレイヤがゲームへの参加も含めてランダムに意思決定をした場合は（３人とも強化学習プレイヤ１）で $\lambda=0$ と同じである。このときの MSD は約 113 であった。

各トリートメントにおいて，MSD の値をスムージングするために MSD を従属変数，λ，η，ω を従属変数として，重回帰分析でモデルを推定した。その推定値を用いて最小 MSD を探索した。

10）λ は 0.2 刻み，η は 1 刻み，ω は 0.5 刻みで計算機実験を行った。

11）$(OC_t, SC_t)=(0, 4)$ または $(0, 4)$ のとき，MSD は最大となる。

1-1 PG first トリートメント

表2 シミュレーションおよび推定結果（PG firstトリートメント）

	min MSD	lambda	eta	omega	estimated min MSD	lambda	eta	omega
0REL	68.71129	7	193	0	80.97906	7.8	190	0
1REL	70.03942	5.8	198.5	0	79.61186	7.8	220	0
2REL	76.08646	5.8	205	0.5	85.42284	7.8	190	0
3REL	82.33933	7.4	210	−	87.53748	7.8	220	−

出典：筆者作成

表2がPG firstトリートメントにおける計算機実験および推定結果である。各条件でλが大きく，η，ωが小さい場合，最小MSDを与える。

これらのMSDの値は均衡MSDよりも明らかに小さい。すなわち，強化学習モデルは均衡予測に比べて，人間行動をより近似していると言える。推定結果を見ると3人グループ中1人だけが強化学習モデルその1に従い，そのほか2人が強化学習モデルその2に従っている場合が最もMSDが小さい。強化学習モデルはランダム意思決定よりも人間行動をより近似している。

パラメータについて解釈しておこう。λが大きいことは行動を早い段階で確定させる傾向が強いことを意味する。ηについては解釈が微妙であるが大きい場合には今期の利潤を余り反映させない方がフィットがよいことを意味し，小さい場合にはその逆のことを意味する。$\omega=0$はヒル・クライミングで意思決定を行うときに直前期の利潤を最も強く反映させることを意味する。

1-2 LC first トリートメント

表3 シミュレーションおよび推定結果（LC firstトリートメント）

	min MSD	lambda	eta	omega	estimated min MSD	lambda	eta	omega
0REL	99.60943	7.8	208.5	2.5	113.0516	7.8	190	3
1REL	91.37542	2.6	205.5	3	112.2583	0	220	3
2REL	93.84268	2.4	209.5	1	112.877	0	190	3
3REL	94.1203	2	195.5	−	111.9536	7.8	190	−

出典：筆者作成

表3はLC firstトリートメントの計算機実験および推定結果である。最小MSDを与えるパラメータは0RELおよび3REL条件ではηが最大値で，それ以外の条件では，ηは最小値である。一方，ωは最大値である。

これらのMSDの値は均衡MSDよりも明らかに大きい。すなわち，均衡予測は強化学習モデルに比べて，人間行動をより近似していると言える。これはPG firstトリートメントと逆の結果である。さらに，最小MSDを与えるパラメータの方向性が一定でないところからも，強化学習モデルがLC firstトリートメントの実験結果をうまく再現できていないことがわかる。

また，ランダム意思決定のMSDが約113であったことをふまえると，この結果は強化学習モデルの説明力がランダム意思決定と高々同じであることを示している。以上から，LC firstトリートメントでは強化学習モデルが実験結果をうまく説明できたとは言えない。

Ⅵ 議 論

計算機実験の結果をまとめよう。PG first トリートメントでは実験結果がうまく説明できた一方，LC first トリートメントでは実験結果を説明できなかった。なぜこのような理由が生じたのか。本節では強化学習とそのバリエーションの説明力がトリートメントが異なると大きく異なった理由について検討する。

手がかりとして囚人のジレンマがペアを変えずに多数回繰り返される場合の実験結果を考えよう。ラウンドが進むに従って，平均協力率が高くなるのが典型的な結果である[12)]。強化学習モデルはこの結果をうまく説明できる。すなわち，初期ラウンドでは試行錯誤がなされるため，協力率が余り高くないが，ラウンドが進むにつれて協力率が増加するのである。これは協力したときに得られる利得が高いので，それを選択する q 値がラウンドを経るごとに大きくなり，協力を選択する確率が高まることを意味している。

この事実をふまえれば，PG first トリートメントで実験結果がうまく説明できた理由が明らかになる。このトリートメントは最初の 30 ラウンドで PG，後半の 30 ラウンドで LC をプレイするものであり，被験者は前半ではうまく協力できず，後半で高い協力を示した。計算機実験との兼ね合いで言うと，試行錯誤が行われる初期ラウンドで PG がプレイされている。この間の仮想プレイヤの協力率は後半に比べて低くなるだろう。一方，後半では協力を均衡として持つ LC がプレイされている。後半にプレイするゲームが PG であったとしても強化学習アルゴリズムを持つ仮想プレイヤは協力しやすい傾向がある。さらに今回の後半ゲームは均衡が協力であるゲームである。より協力しやすい状況になっている。以上から，前半で協力がうまく成立せず，後半で協力が成立する図式がより成り立ちやすいことがわかる。

一方，LC first トリートメントで強化学習およびそのバリエーションが実験結果をうまく説明できなかったのは以下の理由による。強化学習では上述したとおり，初期ラウンドでは行動が定まらず，試行錯誤の意思決定が行われる。その後，ある拠出量の q 値が大きくなり，その拠出量が多くの場合採用されることになる。また，強化学習では傾向的に高拠出を達成する傾向にある。すなわち，強化学習では初期ラウンドでは拠出率が低く，後半ラウンドでは拠出率が高くなる。この結果は被験者実験のそれと逆である。よって，強化学習およびそのバリエーションは被験者の意思決定をうまく再現できなかったのである。

本稿での被験者実験には Knez and Camerer［2000］や Knez［1998］の言う precedent transfer が存在する。この考え方は過去にプレイしたゲームで均衡を選択していると，その結果が引き続いてプレイされるゲームにおける人間行動に影響を与えるという議論である。よって，「precedent transfer が存在する経済実験での人間行動を説明するには，基本的な強化学習モデルでは不十分である」ことが本研究から明らかになった。

Ⅶ 結 論

本稿では，実験の前半と後半でゲームの均衡が異なる経済実験を例にして，強化学習およびその

12）Erev and Roth［1998］の被験者実験と計算機実験の比較を参照せよ。

バリエーションが，被験者行動をどの程度説明するか検討した。その結果，実験の状況によっては説明力が悪くなり，ナッシュ均衡予測よりも悪い場合があることが明らかになった。これまで強化学習モデルはナッシュ均衡予測に比べて，実験結果をより適切に説明すると言われてきた。しかし，ここで示したように，強化学習モデルおよびそのバリエーションではうまく説明できない人間行動も存在する。

これは単純な強化学習モデルの限界を示している。初期ラウンドにおいて，強化学習モデルは被験者ほど賢明でない。初期ラウンドの被験者行動をより適切に説明するモデルを開発するには，やはり強化学習以外の意思決定アルゴリズムも包摂した意思決定モデルを構想する必要があろう。しかし，その際には，どの要因がどのような効果を持っているかを見極められるように，なるべく単純な形のモデルにすることが肝要である。

強化学習以外の意思決定アルゴリズムを実装する方法として，たとえば，強化学習に揺らぎを加えたモデル，複数の意思決定アルゴリズム（仮想プレイ，ナッシュ均衡，ランダム意思決定，ロス回避など）を実装し，それらの選択を強化学習で行うというエレブ―グレッグ型の意思決定モデル（RELACS, Erev and Greg［2005］）などが考えられる。これらの意思決定モデルでも実験の前半と後半で均衡が変化する状況でどの程度被験者行動を説明できるかどうかは検討されていない。また，計算機実験では1ラウンド目の選択確率（初期確率）は一様分布で与えられる。そこで初期確率を被験者実験の第1ラウンドで実際に見られた選択結果に変更することで，その後の学習プロセスが変化するかもしれない（Iwasaki, et al.［2003］）。これらの検討は今後の課題としたい。

参考文献

Camerer, C. F.［2003］*Behavioral Game Theory*, Princeton University Press.

Erev, I. and Roth, A.［1998］"Predicting How People Play Games : Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," *American Economic Review*, 88(4), pp. 848–81.

Erev, I. and Roth, A.［2001］"Simple Reinforcement Learning Model and Reciprocation in the Prisoner's Dilemma Game," *Bounded Rationality : The Adaptive Toolbox*, pp. 215–231.

Erev I. and Greg Barron［2005］, "On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies." *Psychological Review*, 112, pp. 912–931.

Fehr, E. and Klaus M. Schmidt［1999］, "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics* vol. 114(3), pp. 817–868.

Güth, W., R. Schmittberger, and B. Schwarze［1982］, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, vol. 3(4), pp. 367–388.

Fischbacher, U.（2007）"z-Tree : Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics*, vol. 10(2), pp. 171–178.

Iwasaki, A., S. Imura, S. H. Oda and K. Ueda［2003］, "Accidental" Initial Outcome and Learning in Experimental Games with Multiple Equilibria, the paper presented at Economic Science Association International Meeting, Available at http://www.cc.kyoto-su.ac.jp/~oda/PDF_files/Iwasaki2003.pdf.（2009.10.6確認）

Iwasaki, A., K. Ogawa, M. Yokoo and S. H. Oda［2007］"Reinforcement Learning on Monopolistic Intermediary Games : Subject Experiments and Simulation," in *Agent-Based Approaches in Economic and Social Complex Systems IV*

Kahneman, D., J. Knetsch, and R. Thaler［1986］"Fairness as a Constraint on Profit Seeking : Entitlements in the Market," *American Economic Review*, vol. 76(4), pp. 728–41.

Kahneman, D. and A. Tversky［1979］, “Prospect Theory : An Analysis of Decision Under Risk,” *Econometrica*, Vol. 47, pp. 263–292.

Knez, M., 1998, “Precedent Transfer in Experimental Conflict-interest games,” *Journal of Economic Behavior & Organization*, vol. 34, pp. 239–249.

Knez, M. and C . Camerer［2000］, “Increasing Cooperation in Prisoner’s Dilemmas by Establishing a Precedent of Efficiency in Coordination Games,” *Organizational Behavior & Human Decision Processes*, vol. 82（2）, pp. 194–216.

Ledyard, J.［1995］, “Public goods experiments,” *Handbook of Experimental Economics* edited by J. K. Kagel and A. Roth. Princeton : Princeton University Press, pp. 111–194.

McKelvey, R. D. and T. R. Palfrey［1995］, “Quantal Response Equilibria for Normal Form Games,” *Games and Economic Behavior*, vol. 10（1）, pp. 6–38.

Roth, A. and I. Erev［1995］, “Learning in Extensive-Form Games : Experimental Data and Simple Dynamic Models in the Intermediate Term,” *Games and Economic Behavior*, vol. 8（1）, pp. 164–212.

Selten, R.,［1998］, “Axiomatic Characterization of the Quadratic Scoring Rule,” *Experimental Economics*, vol. 1, pp. 43–62.

Simth, V. L.,［1962］, “An experimental study of competitive market behavior,” *The Journal of Political Economy*.

Stahl, Dale O.［1998］, “Is step-j thinking an arbitrary modeling restriction or a fact of human nature ?,” *Journal of Economic Behavior & Organization*, Vol. 37, pp. 33–51.

Stahl, Dale O. and Ernan Haruvy, in press, “Level-n bounded rationality in two-player twostage Games,” *Journal of Economic Behavior & Organization*.

Tversky, A. and D. Kahneman［1981］, “The Framing of Decisions and the Psychology of Choice,” *Science*, vol. 211（30）, pp. 453–458.

Tversky, A. and D. Kahneman［1986］, “Rational Choice and the Framing of Decisions,” *Journal of Business*, vol. 59, 4, S251–S278.

重田正美［2005］,「地域通貨の将来像―スイスの地域通貨「WIR」の事例を参考に」,『調査と情報―ISSUE BRIEF』, 第 484 巻, 1–10 頁.

二村英夫, 小川一仁, 高橋広雅［2009］,「地域通貨の使用体験が公共財供給にもたらす影響―経済実験による考察」, 広島市立大学国際学部ワーキングペーパー。

三浦一輝［2008］,「「地域通貨」の流通に関する理論分析」『法政大学大学院紀要』第 60 巻, 69–76 ページ。

室田武［2004］,『地域・並行通貨の経済学―一国一通貨制を超えて』, 東洋経済新報社, 東京。

西部忠［2002］,『地域通貨を知ろう』岩波書店, 東京。