人工市場とマルチエージェント強化学習

論　文

人工市場とマルチエージェント強化学習

谷　田　則　幸*

概　要

現在の経済学の多くの分野におけるアプローチには、数学的方法が主として用いられる。経済（現象）という複雑なものをモデル化し、議論を簡単にするためには数学という道具が不可欠であるといえよう。しかし、モデル化により多くの情報を取りこぼす結果になることも事実である。

本稿では、経済学における数学的手法によるアプローチの問題点を探り、人工市場の存在意義とその可能性について考察する。また、人工市場の実現を技術的に支えるコンピュータサイエンス的手法としてのマルチエージェント強化学習をによるアプローチを概観し、報酬配分問題で起こる非合理的ルール抑制に関して従来手法に改良を与える。

キーワード：人工市場；機械学習；マルチエージェント理論；強化学習
経済学文献季報分類番号：02-21

1．はじめに

株式市場や外国為替市場に代表される金融市場は、近年極めて激しい変動を呈している。いうまでもなく、そういった市場は市場参加者である個々の人間が持つ見通し、賭け、騙し、かけ引きといった複雑で、非常に人間臭い要素が絡み合って形成されている。ところが、伝統的な経済理論においては、合理的な人間のみが存在し、複雑さを排除するために過度に理想的で、ある意味で非現実的な市場を想定している。こういった現実市場と理論市場に存在する隔たりを説明するためには、市場参加者の個々の特性や心理的側面を無視するわけにはいかない。

古典派経済学から新古典派経済学への転機には、限界革命と数学の導入が主要な役割を果たした。数学モデルの典型として一般均衡理論があるが、そこで用いられる仮定には現実市場になじまないものもある。たとえば、一般的な市場は常に価格変動があり、消費者は効用を最大にしているとはいいがたい。ところが、経済理論においては、価格体系が変動しない

＊ E-mail：tanida@ics.uci.edu

とか、個々の経済主体はその価格体系の下で目的関数を最大化している、という仮定がなされている。もちろん、これらの仮定は数学的アプローチを推し進める上で必要な仮定ではある（[西村98]、[塩沢00b]）。

自然科学・工学あるいは広く科学全般における研究様式の重要なものとして「実験」があげられる。「実験」の対極に位置するものとして「理論」があげられる。簡単には、「実験」により「理論」を形成し、形成された「理論」を「実験」により検証する。両者は、相互にフィードバックしながらより完成度の高いものになる。

しかしながら、実験にはさまざまな意味で限界が存在することは容易に想像できる。たとえば、多くのデータを必要としたり、対象が非常に巨大であったり、一回の実験にかかる費用が莫大なものとなるなどの理由が考えられる。このような場合には、「実験」に代わって「コンピュータ実験」いわゆるコンピュータ・シミュレーションが主流になりつつある。また、古典的な数学的アプローチでは解き方は分かっていても、実際に解くのが難しかったり、解くのに多くの時間を必要としたりする場合がある。そのような場合にもシミュレーションは威力を発揮する。

一方、経済学を含む社会科学に目を転じると、小規模の実験はあるものの、実験がほとんど不可能か、極めて困難なものとされてきた。その理由は明らかで、「人間社会」という実際社会があるからである。そのような社会の存在に影響を及ぼすような実験は許されるものではない。したがって、ここでもシミュレーションが大きな役割を持ち得る。ここでいうシミュレーションは、経済学においてよく利用されるモデルを固定した状態でのパラメータ調整によるシミュレーションのことではなく、後述のエージェントベースシミュレーションを意味する。

本稿では、市場参加者の心理的側面、数学的理論で説明しがたい問題をエージェント理論による人工市場について考察する。本稿であつかう人工市場（広くは、人工社会）は、複雑性を持った主体の行動をエージェントとしてモデル化するエージェント・ベースト・モデリング（単に、エージェントモデリングともいう）とコンピュータシミュレーションによってエージェントのミクロ・マクロな挙動を解析するエージェントベースシミュレーションという形で研究されている。また、特にエージェントの学習能力、学習効率とエージェントが機能する上で必要となる内部複雑性についても議論する。

2．エージェントモデルと社会科学

エージェントという言葉は、かなり以前から社会科学分野でも用いられている。産業組織論、経営組織論などでのプリンシパル・エージェント理論がその代表であろう。また、ゲー

ム理論で言うところのプレーヤーとして考えることも出来よう。プリンシパル・エージェント理論では、プリンシパル（依頼人）とエージェント（代理人）間に存在する依存関係を表し、そこで起きる問題解決を提唱している。たとえば、株主と経営者、上司と部下、サービスの代理人と委託者といった関係がこれにあたる。これらの依存関係は、ある面で上下関係を表しているといえるが、持てる情報の非対称性も現わしている。すなわち、代理人が依頼人よりも多くの情報を有しているということであり、また、情報を持たないものが情報を持つものをコントロールしなければならない、ということを意味している［HT99］。

一般的に、エージェントは自律的（自身の基準・判断により行動を選択する）で、自己の目的にかなった行動をとる。エージェントが複数存在する場合には、エージェントの集合体という意味でエージェンシー［Minsky52］とかマルチエージェントと呼ばれる。マルチエージェントモデルでは、単にエージェントが複数存在するというだけでなく、エージェント間の交互作用というものを考える必要がある。

3．人工市場研究

エージェントモデルによる人工市場は、3つの要素（フェーズ）で形式化される。第一は、市場参加者としてのエージェントである。エージェントは、自立的で、複数存在し、経済的行動（投資、交換など）を行いながら学習するという行動を繰り返す。一般にエージェントは、プログラムで実現され、何らかの情報を入力し、何らかのルールに基づき行動を決定し、行動を出力する、という Input-Output システムであり、その構造は時間とともに変化する。第二は、市場構造のモデルであり、エージェントの経済行動の結果に基づく価格決定メカニズム（相対型・均衡型）をさす。第三は、先の2つのフェーズによるシミュレーション結果の解析により、ミクロ的な行動とマクロ的現象との関係の考察である。

［和泉00］は、人工市場研究をその目的別に以下のように3つに大別している。

（1）経済史派

貨幣の創発など大きな経済制度の変遷や経済史の説明

（2）実験ツール派

人間や計算機の取引プログラムを参加させた仮想取引実験の為のシミュレータ構築

（3）市場分析派

バブル崩壊などの金融市場に見られる個別の経済現象や市場メカニズムの解析

この中では、（2）と（3）のアプローチが主流であり、コンピュータサイエンスが関わるのもこの二つである。（2）においては、共通テストベッドとしてのバーチャル市場という位置付けで、U-Mart と呼ばれるシミュレータが開発されている（［塩沢00a］、［川村他00］など）。

経済学の側面からは、株式など売買行動に関する人間の判断様式の解明や、市場の乱高下などの市場の投機的な動きを回避するための実験をそのシミュレータを使って行うことを目的としている。(3)には、「複雑系」で有名なサンタフェ研究所のArthurらの人工株式市場がある［Arthur91, Arthur95］。Arthurらの人工株式市場は、100個オーダー程度の条件式（IF-THEN）からなるルール集合を持つエージェントを定義して、シミュレートによる価格と合理的期待仮説に基づく価格を比較し、おおよその一致を見たが、ファンダメンタルズによる価格から著しく乖離したり、接近したりと、同仮説では説明できない現象を示した。もちろん、個々のエージェントが持つ条件式（ルール）は動的に変更される（エージェントの学習）が、これには遺伝的アルゴリズム（GA：Genetic Algorithm）（［Holland92］、［HM91］など）が用いられている。また、個々のエージェントの学習に遺伝的プログラミング（GP: Genetic Programming）［Koza92, Koza94］を用いた人工株式市場のシミュレーションがChenらによって開発され、合理的期待仮説の検証、効率的市場仮説の検証、価格変動の複雑度の現実市場との比較などがある（［CY96］、［CK99］、［CT99］、［CY99］、［CYL99a］や［CYL99b］など）。

ArthurやChenの結果からも分かるように、エージェントアプローチにはエージェント自身の学習が不可欠である。もし学習能力を持たなければ、静的なエキスパートシステムの分散化に過ぎない。また、この学習能力の良し悪しがエージェントがなす行動の的確さを左右し、学習効率がエージェント（あるいはマルチエージェント）がゴール（解）へ至る速さに影響することは言うまでもない。いわば、閉じた系ではなく開かれた系を求めているといえよう。

4．エージェントと学習

4．1　強化学習の枠組み

強化学習（reinforcement learning）は、試行錯誤をしながら学習し、徐々に環境に適応する学習の枠組みを指す。強化学習に対する枠組みとして、教師付き学習（supervised learning）がある。教師付き学習には、文字通り教師の存在が仮定されており、入力として状態を与えると、出力として行動が明示されるという、ある種の神託（オラクル）がある。一方、強化学習には教師の代わりに報酬（reward）が定義される。この報酬は、サラリーマンの年俸制をイメージすれば理解しやすいであろう。年俸制では、ある一定の目標が設定され、それをクリアすればその年俸は保証され、大きくクリアすれば臨時ボーナスが支給されたり、次期年俸査定に好影響を与える。あるいは、年俸の大小によりモラルが上がったり、下がったりするかもしれない。強化学習における報酬は、この設定目標とその代価を表

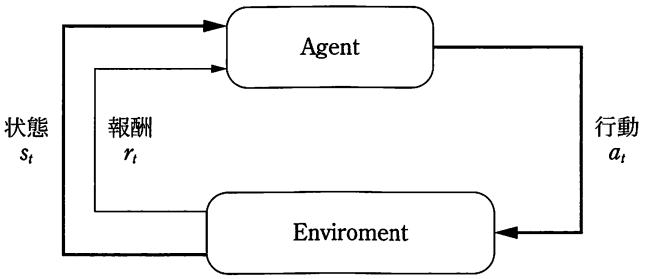

図１：強化学習の枠組み

Step. 1	時刻 t にエージェントは環境から状態を観測（状態観測）して s_t を得（入力）、それをもとに意思決定をし行動 a_t をとる（出力）。
Step. 2	エージェントの行動により、環境の状態は s_{t+1} に変化（遷移）する。 また、その遷移に対する報酬 r_t がエージェントに与えられる。
以降、時刻 $t+1, t+2, \cdots$ においてゴールに達するまでこれらを繰り返す。	

図２：強化学習のアルゴリズム（概要）

わしている。また、強化学習では設定目標が達成されなかったときのための罰（ペナルティ）も定義される。一般的なサラリーマンの給与体系では、平均的な仕事をしている限りにおいてその所得は保証されており、ほどほどに仕事をするという人間が現われるかもしれない。そういった人間の動機付けとして、すなわちズルを防ぐための方策として考えればいいかもしれない。さらに、報酬には遅れも加味される（delayed reward）。これに対し、通常の報酬を即時報酬（immediate reward）ということもある。強化学習の枠組みを図１に示す。

強化学習には、学習主体としてのエージェントと制御対象としての環境が定義される。環境を制御するのはエージェントなのでエージェントはコントローラとしての働きを持つ。また、強化学習は制御学習（Control Learning）の一種であり、ロボティクス分野での研究も多い。

強化学習の一般的なアルゴリズムを示す（図２）。基本的に、エージェントは利得（return）が最大になることを目指し、その基準のもとで何をするべきか（どう行動すべきか）を学習する。また、エージェントは事前に環境に関する知識をもたない（エージェントに環境の事前知識をもたすことは、実装では制御プログラムのコーディングを意味し、設定者に多大な負荷がかかることを意味する）。大雑把に言えば、設計者はやらせたい仕事「何をすべきか」をエージェントに与え、ゴールに達すれば報酬を与える、ということを指示するだけで、ゴールへの到達方法についてはエージェントの試行錯誤により自動的に「どうやって仕事を片付けるか」が獲得されるということである。もう一つの特徴として、強化学習は報酬の遅れを考慮にいれており、一つ一つの準備的行動で報酬は得られないかもしれないが、それが

重なってゴールに到達すれば大きな報酬が得られる、という現実社会でありそうなことも表現できる。

4．2　強化学習

強化学習の歴史は、Minskyの初期の論文にまで遡ることができる［Minsky52］。この分野での研究には、現在に至るまで時間的相違（Temporal Difference）に基づくBellmanのDynamic Programming［Bellman57］、Samuelの対戦ゲーム（チェス）［Samuel59, Samuel67］などがある。概要を知るためには、［SB98］が詳しい。試行錯誤によるものとしては、WatkinsのQ-Learning［Watkins92］があり、現在の強化学習研究の主流のひとつとなっている。

また、強化学習は試行錯誤により学習をするので、専門家が得た解よりも優れた解を発見することが強く期待され、不確実性・複雑性を持つ対象に対するアプローチにも可能性を秘めている。

次に、強化学習を形式的に記述するためのモデルを考える。

4．2．1　マルコフ決定過程

まず、環境のダイナミクスをモデル化するためにマルコフ決定過程（MDP：Markov Decision Process）を考える（図3）。MDPは（S, A, p, R）の4項組で次のように定義される。

いま、環境の取る状態の集合を、$S=\{s_1, s_2, \cdots, s_n\}$、エージェントの取る行動の集合を$A=\{a_1, a_2, \cdots, a_n\}$とする。状態$s$において行動$a$を起こしたときに状態が$s'$に遷移したとする。このとき、状態遷移関数は以下のように記述される。また、このときエージェントに確率的に与えられる実数値スカラー量を即時報酬といい、その期待値を$R_a(s, s')=R_a(s)$で表わす（状態s'には依存しないため、右辺のように書けることに注意しよう）。

$$\delta_p(s, a)=s' \quad (s\in S, a\in A, s'\in S)$$

ただし、$p=P_a(s, s')=\Pr\{s_{t+1}=s' \mid s_t=s, a_t=a\}$はそのときの状態遷移確率をあらわす。

また、状態および報酬にはマルコフ性（この場合は一次マルコフ性）、すなわち、

$$\forall q\in\{s, r\}:\Pr(q_{t+1} \mid q_t, q_{t-1}, \cdots)=\Pr(q_{t+1} \mid q_t)$$

が仮定されており、それがマルコフ決定過程という名前の由来である。

さらに、エージェントが状態sにおいて行動aを選択することを政策（policy）πを取ると

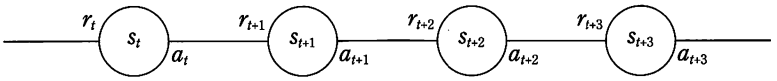

図3：マルコフ決定過程

いう。πは次のような関数として定義される。その集合Πを政策集合と呼ぶことにする。

$$\pi : S \to A$$

環境がMDPで定義されているとする。エージェントは観測入力として環境の状態をエージェント自身が入手する。いま仮にエージェントの観測能力が低く、本来sのところを$\tilde{s}$と観測してしまった場合、状態遷移確率が定義できず、MDPが完全に定義されない。このような場合には、部分観測マルコフ決定過程（POMDP：Partially Observable MDP）として定義される。これは、不完全知覚問題として扱われる。

また、報酬の遅れが存在するときには割引率（Discount Factor）$\gamma \in [0, 1]$を考慮する。強化学習では、これを用いた利得を以下のように定義することも多い。$\gamma=1$のときは、単純に報酬の和になることに注意しよう。

$$V_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

MDPにおいてエージェントが定常政策πを取る場合には、利得の期待値は状態sだけに依存し、時間tには依存しないので、とくにState-Value関数といい、$V^{\pi}(s)$と書く。また、すべての状態sにおいて次の性質をみたすとき、政策πは政策π'よりも優れているという。

$$V^{\pi}(s) \geq V^{\pi'}(s) \quad \text{for } \forall s \in S$$

さらに、すべての$\pi' \in \Pi$に対して上の関係が成り立つとき、すなわち、

$$\exists \pi^* \in \Pi \text{ such that } V_{\infty}^{\pi^*}(s) = \max_{\pi \in \Pi} V^{\pi}(s) \text{ for } \forall s \in S$$

のとき、π^*は最適政策であるといわれる。最適政策は必ずしもユニークではないことに注意しよう。また、最適政策を求めることこそエージェントの目的であることを思い出そう。

同様に、状態と行動の対(s, a)の関数をAction-Value関数と呼び、$Q(s, a)$と書く。また、その値をQ値と呼ぶ。関数Qは状態sにおいて行動aを取った後、政策πを取りつづけるときの利得の期待値を表わしている。そのため、明示的に$Q^{\pi}(s, a)$と書くこともある。

最適政策とValue関数に関してまとめたのが以下の定理である。

【定理1】[SB98]　有限なMDPにおいて以下のような最適Value関数がユニークに存在し、Bellmanの最適方程式に従う。

$$V^*(s)=\max_a Q^*(s,a)=\max_a R_s^a+\gamma\sum_{s'}P_a(s,s')V^*(s')$$

$$Q^*(s,a)=R_s^a+\gamma\sum_{s'}P_a(s,s')\max_{a'}Q^*(s',a')$$ ■

このとき、ダイナミックプログラミングを用いて次の系を導くことができる。

【系1】　状態遷移確率 $P_a(s,s')$ と即時報酬 $R_a(s)$ が既知ならば、DPにより【定理1】中の式をみたす最適解（最適政策）を求めることができる。 ■

しかし、【系1】には2つの問題点がある。第1の問題点は、$P_a(s,s')$ と $R_a(s)$ が既知であるという過程である。これが既知であるということは、事前に環境のモデルの一部を知ることになるが、一般的にはそれは難しいことである。第2の問題点は、事前にそれらを知っていたとしても、【定理1】の方程式を解くシステムは一般的にとてもおそいことが分かる。なぜなら、ある状態に隣接するすべての集合上の和を取っているからである（両式最右辺第2項）。

4.2.2　*Q*-Learning

前小節で指摘した問題点を克服する一つの解として、【系1】で必要とされた前提知識を仮定せずに適切な政策を得ることを考える。ここでは、そのための最も一般的なアルゴリズム *Q*-Learning［Watkins92］を示す。

Q-Learningは、エージェントが環境との間で試行錯誤による相互的作用を繰り返しながら最大の *Q* 値 $Q^*(s,a)$ を推定するアルゴリズムで、Watkinsにより提案された（図4）。ここで、図4中の α, γ $(0\leq\alpha,\gamma\leq1)$ はそれぞれ学習率、割引率を表わしている。また、手続き4で分かるように、遷移先 s' の最大の *Q* 値に基づいて遷移元の *Q* 値を推定していることが分かる（ブートストラップ法）。

Q-Learningの最大の利点は、エージェントの多くの行動の繰り返しにより、最適な *Q* 値を得られることが保証されていることである。

【定理2】（*Q*-Learningの収束定理）［SB98］

行動選択で、エージェントがすべての行動を十分多く選択し、かつ学習率 $\alpha(0\leq\alpha\leq1)$

初期化：Q 値のテーブルを初期化する；
　　For $\forall s \in S, \forall a \in A$　　$Q(s,a)=0$　　（あるいは、小さな確率値でもよい）
　　初期状態を s_0 とする
手続き（時刻 t）：以下を繰り返す；
　　1．環境の状態 s を観測する（ただし、時刻 0 のときは s_0）
　　2．行動政策（行動選択法、戦略）にしたがって、行動 a を実行する
　　3．報酬 r を受け取り、次の状態 s' を観測する
　　4．次の更新式により Q 値を更新する

$$Q(s,a) \leftarrow Q(s,a) + \alpha \{ r + \gamma \max_{a'} Q(s',a') - Q(s,a) \}$$

　　5．時刻ステップを $t+1$ に進めて、1 へもどる

図 4 ：　*Q*-Learning アルゴリズム

が時間に関して適切に減少する（すなわち、$\sum_{t=0}^{\infty} \alpha(t) \rightarrow \infty$ かつ $\sum_{t=0}^{\infty} \alpha(t)^2 < \infty$ をみたす時間 t の関数となっている）とき、Q-Learning アルゴリズムが得る Q 値は最適な Q 値に確率 1 で収束する。 ■

【定理 2】で分かるように、Q-Learning アルゴリズムはある条件を満たすかぎり、有用なアルゴリズムといえる。収束そのものは政策とは独立であり、極端には確率的な政策でさえ収束する。また、そのシンプルさから実装が容易ある。一方で、Q-Learning もダイナミックプログラミング手法によるので、時間的効率は良いとはいえない。一般的な貪欲（greedy）戦略のほかに、確率的に行動を選択する Boltzmann 探索法も考えられている。

また、部分観測マルコフモデルのような場合で、二つの状態を区別できず、かつそのそれぞれの状態に対する最適行動が一致しないような場合には、Q 値の推定に影響を及ぼすことが指摘されている［宮崎他99a, 荒井01］。

Q-Learning と同じダイナミックプログラミング手法による強化学習アルゴリズムとして Monte-Carlo（MC）法に基づくものがある。Q-Learning は遷移先の観測の最大 Q 値に基づいて Q 値を推定するのに対し、MC 法では実際の報酬によって Q 値を推定しているのが大きな相違点である。したがって、MC 法は非ブートストラップ法といえる［Sutton98］。

4．2．3　Profit-Sharing アルゴリズム

エージェント強化学習でよく用いられる別のアルゴリズムとして、Profit Sharing アルゴリズムがある。Q-Learning などのダイナミックプログラミング系の強化学習アルゴリズムは、最終的にできるだけ多くの（最大の）報酬をを得ることを目指している（最適性の追求）。一方、Profit Sharing は学習の途中段階でもコンスタントに報酬を得続けることを目指している（効率性の追求）。特徴として、前者は時間効率は悪いが MDP 環境下では最適解

を得ることができ、後者は時間効率は良いが最適解を得られるとは限らない、ということがある。

Profit Sharingを簡単に説明しておく。

有限（たとえば、T）ステップの状態、行動、報酬からなる系列をエピソードという。$a_t \in A, s_t \in S, r_t \in R (t=0, 1, ..., T)$ に対して、

$$s_0, a_0, r_1, s_1, a_1, r_2, ..., a_{T-1}, r_T, s_T$$

Profit Sharingは、エピソード中に含まれる状態-行動対（s_t, a_t）の各々について、次のような強化を一括して行う。

$$W(s_t, a_t) \leftarrow W(s_t, a_t) + f(t, r_T, T)$$

ここで、Wは状態-行動対の価値（ウェイト）を表わす関数、fは強化関数である。

4．3　マルチエージェント強化学習

ここでは、マルチエージェントに対する強化学習を考える。前述の（シングルエージェントに対する）強化学習の枠組みをマルチエージェント化することにより発生する問題を考えてみよう。

4．3．1　マルチエージェント化による問題

一般的なマルチエージェント強化学習の枠組みを示す（図5）。シングルエージェント（図1）との違いは、複数のエージェントが存在し、一つの集合体であるエージェント群が形成されていることである。ここでは、報酬がエージェント群に与えられていることに注意されたい。すなわち、個々のエージェントにではなく、エージェント群（言い換えれば、システムの全体）に目標が設定されている。もちろん、個々のエージェントに直接設定することは可能ではあるが、モデルの設計が難しくなってしまう。したがってマルチエージェント強化学習では、図5のようにするのが一般的である。このとき、個々のエージェントの働きに応じて報酬が分配されなければならず、どのエージェントにいくら報酬を与えるかということが問題となる。これを報酬配分問題といい、第1の問題である。[荒井01]は、他のエージェントを環境の一部とみなす単純なアプローチには非ブートラップ法のMC法やProfit Sharingがよく、報酬がその都度与えれるような場合には、計算量的にはProfit Sharingの有望性を指摘している。

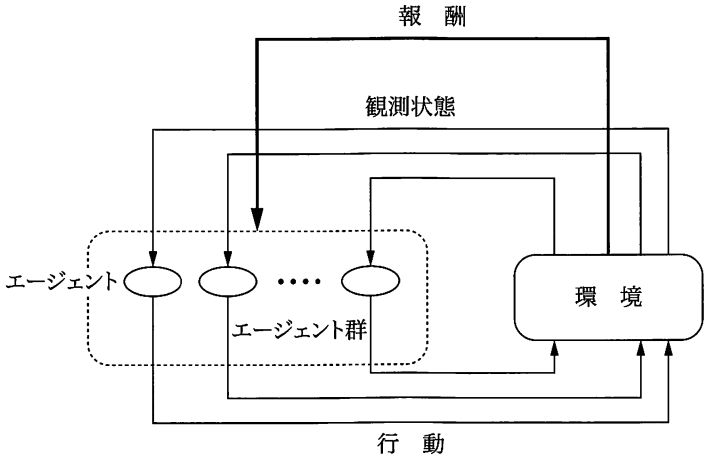

図5：マルチエージェント強化学習の枠組み

【例1】　アメリカンフットボールは、11人のプレーヤーが攻撃と守備に別れ相手陣地にボールを持ちこんで得点するスポーツであり、非常に戦略が重要となる。ボールの持ち込み方にはランとパスがあり、中でもショットガンフォーメーションは極めて華麗である。ショットガンでは、パサーとレシーバーが注目されるが、実際には他の多くの味方選手がそれを成功させるための役割を果たしている。ボールを出す選手、パサーが確実にパスできるように相手の攻撃を防ぐ選手、フェイントをかける選手、フェイクのレシーバーなどなど。まさに、チームプレイで一つの仕事を成し遂げている。エージェント学習でも、そのような「縁の下の力持ち」エージェントがいる。それらにもきちんと仕事に見合った報酬を与えなければ、正確な学習は継続されない。　■

第2は、前述の不完全知覚問題である。シングルエージェントの場合にも起こるこの問題は、マルチエージェントの場合はさらに深刻である。なぜなら、エージェントが知覚すべき状態空間が明らかに複雑で、巨大化することが予想されるからである。

第3の問題は、同時学習問題と呼ばれる。エージェントは互いに協調あるいは競争しながら、すなわち関連しながらそれぞれの役割を果たす。もし、個々のエージェントが取る政策が定常であるならば、自己（一つのエージェント）が取るべき行動、遷移する状態を把握できる（しやすい）であろう。しかし、定常でない場合、あるいは定常でない政策空間を持つエージェントが多くいるような場合には話がややこしくなる。すなわち、自己の遷移する状態は、自分自身の行動のみによるのではなく、他のエージェントの影響を受けたり、あるいは連携して行動した結果による。例えれば、多変数関数を一変数関数で近似するようなもの

である。したがって、ブートストラップ法よりも非ブートストラップ法のほうがこのような問題には向いているといえよう。ただし、不完全知覚問題がない状況下での追跡問題で、ある一定の条件が見たされれば、Q-Learningでも最終的には最適解を得ることが示されたことは、非常に興味深い［荒井他98］。

4．3．2　報酬配分問題に対する考察

前小節で見たように、マルチエージェント学習における問題点を考えると、対象には依存するが、Profit Sharingが平均的に良い性能を示しているといえる［荒井98］。そこで、本稿を結ぶにあたり、報酬配分問題にProfit Sharingを適用した際の非合理的ルールの抑制について議論する。

マルチエージェント強化学習において、直接貢献へ至ることのないルールを非合理的ルールと呼ぶ。全ルールがこれに収束すると、報酬が0となってしまう。貢献したにもかかわらず、報酬が得られないという状況を避けなければならない。これを非合理的ルール抑制という。

非合理的ルール抑制に関する理論的考察が宮崎らによって行われている［宮崎他99b, 保知他02］。いま、報酬Rが発生したときに、報酬発生条件内に示されたエージェントのすべてに直接報酬R、それ以外のエージェントすべてに間接報酬$\mu R(0\leq\mu\leq 1)$が与えられるものとする。

【定理3】［宮崎他99b］　システム全体の単位行動あたりの期待獲得報酬が正となるのは、次の条件式をみたすときであり、かつそのときに限る。

$$\mu < \frac{M-1}{M^{W}\left\{1-\left(\frac{1}{M}\right)^{W_0}\right\}} \cdot \frac{1}{(n-1)L}$$

ここで、nは全エージェントの数、Mは行動の種類の数、Wは直接貢献したエージェントの最大エピソード長、W_0は直接貢献したエージェント以外のエージェントの強化区間、Lは同一感覚入力下に存在する有効ルールの最大競合数である。　■

また、［保知他02］では直接貢献エージェント数xを考慮し、μに関する条件式を次のように提案し、［宮崎他99b］のものより、学習が早くなり、かつ非合理的ルールをよく抑制していることを示している。直接貢献エージェントに関する定義が両者で異なることに注意しよう。以下では、［保知他02］における定義を広義直接貢献エージェントと呼ぶことにす

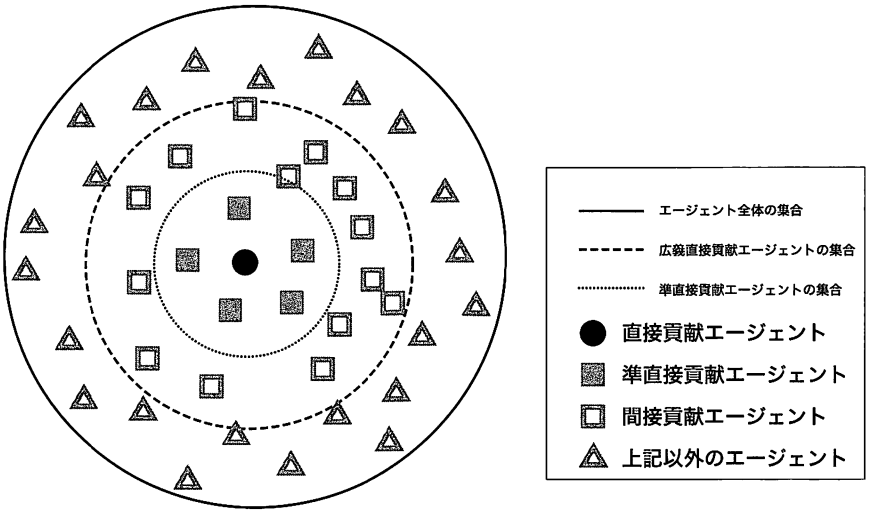

図6：エージェントの種類と関係

る。

$$\mu < \frac{M-1}{M^W\left\{1-\left(\frac{1}{M}\right)^W\right\}} \cdot \frac{{}_{n-1}C_{x-1}}{{}_{n-1}C_x L}$$

いま、エージェント a_i, a_j 間に距離 $d(a_i, a_j)$ を次のように導入する。直接貢献したエージェントの数を x とすると、$n-x$ 個のエージェントは直接貢献しなかったことになる。直接貢献のエージェントの集合を Ψ、直接貢献しなかったエージェントの集合を $\overline{\Psi}(=\Omega-\Psi)$ とする。ただし、Ω は全体集合とする。

（1）　$\forall a, a' \in \Psi$ に対して、$d(a, a')=0$

（2）　$\forall a, a' \in \Omega$ に対して、$0 \leq d(a, a') < \infty$

つぎに、準直接貢献エージェント集合 Φ を次で定義する。ただし、ε は非負。

$$\Phi = \{a \notin \Psi \mid d(a, a') \leq \varepsilon, a' \in \Psi\}$$

エージェントが直接報酬 R を得るとき、強化関数として、行動のバリエーションが M 個の場合のシングルエージェントの場合の合理性定理［宮崎他94］をみたす次のような関数を用いる。

$$f_n = \frac{1}{M} f_{n-1}, (n=1, 2, ..., W_a-1)$$

本稿でもこれを採用する。ここで、f_i は報酬から i ステップ前の強化値、W_a は強化区間を表わす。このとき、初期報酬値 f_0 および強化区間 W_a は次のようになる。ただし，オリジナル

には準貢献エージェントとそれ以外のエージェントの区別がないことに注意しよう。

- 直接貢献のエージェント：　$f_0=R, W_a=W$
- 準直接貢献のエージェント：$f_0=\mu_1 R(\mu_1 \geq 0), W_a=W_1(W_1 \leq W)$
- それ以外のエージェント：　$f_0=\mu_2 R(\mu_1 \geq \mu_2 \geq 0), W_a=W_2(W_2 \leq W)$

いま、$|\Phi|=k$ とすると、次のような必要十分条件を得る。

$$\mu_1+\mu_2 < \frac{M-1}{M^W\left\{1-\left(\frac{1}{M}\right)^W\right\}} \cdot \frac{k+1}{\{n-(k+1)\}L} \qquad (*)$$

証明は、［定理3］に対する［宮崎99b］の証明と同様に、抑制が最も困難な状況で考える。また、基本的な部分は［保知他02］と似ている。

いま、困難な状況においても、非合理的ルールに収束しないように次のように $\mu_i(i=1,2)$ を制限する。すなわち、各エージェントがある1つの非合理的ルールだけを選択して、${}_{n-1}C_{k-1}L$ エピソードで準直接貢献により報酬を得たときの強化値の合計と間接貢献により報酬を得たときの強化値の合計との和よりも、${}_{n-1}C_{k-2}$ エピソードで間接報酬を得た場合の最初に選ばれたルールの強化値が上回るようにする。これを式で表わすと次のようになる。

$$\frac{R}{M^{W-1}} > (\mu_1+\mu_2)R\frac{M}{M-1}\left\{1-\left(\frac{1}{M}\right)^W\right\}\frac{{}_{n-1}C_{k+1}L}{{}_{n-1}C_k}$$

これを変形すると（＊）式が得られる。

$\mu=\mu_1+\mu_2, x=k+1$ としたとき、［保知他02］と等価になることに注意しよう。

よって、準直接貢献エージェントの数が広義直接貢献エージェントの数より少ない場合には、［保知他02］よりも良い性能を示すことが分かる（$\mu=\mu_1+\mu_2$ のとき）。

また、何を準直接貢献とするかだが、たとえばゴールに至る数ステップ前までに関わっていたものなどが自然に考えられる。

5．まとめ

［EA96］には、「エージェントベースのモデリングはトーマス・ロバート・マルサスの時代でも可能であった」という記述がある。人工社会の例としてよく使われるSugarscape程度のものであれば、確かにそのとおりかもしれない。しかし、一方でほんの少し対象が複雑になっただけで、設計が困難になるという事実もある。金融市場のように特別な性格を持つ対象にはシミュレーションにより均衡理論では扱えなかった現象を扱えるようになったという事例もあるが、経済学から見ればまだほんの一部に過ぎない。とても既存の経済学にとって変わるような状況にはない。エージェントベースモデリング、人工市場研究はまだ発展途

上にある。同じことは、複雑系研究についても言えるだろう。これらが光明を見出すためには、マルチエージェント理論の発展、特に強化学習、エージェント間コミュニケーションにおける研究の発展が望まれる。

ただ、従来社会科学ではあまり実施されてこなかった実験をシミュレーションにより実現することには大きな意味がある。それは現在のように未発達な状況においても、エージェントベースシミュレーションの経済学への適用は有効であると考えられる。数学モデルでは確認できないことを「発見」することが大いに期待される。一つの発見のために多くのゴミも排出することにはなるだろうが、実験というのはそういう存在だと思う。

今後は、同時学習問題における Stochastic Game の利用や、機械学習、特に帰納学習との融合などが存在する問題解決の糸口として期待される。また、エージェント理論は従来からの制御、ロボティクスなどの分野のほかに、バイオインフォマティクス分野の研究での研究が期待されている［谷田01］。

謝　辞

本稿は、平成11・12年度関西大学学術研究助成基金（共同研究）ならびに平成 13年度関西大学在外研究の成果報告の一部である。関西大学ならびに共同研究代表研究者・廣江満郎関西大学教授、およびカリフォルニア大学アーバイン校（UCI）ならびに Dennis Kibler UCI 教授に感謝する。

参考文献

［Arthur91］ Arthur, W., Designing Economic Agtents that Act Like Human Agents: A Behavioral Approach to Bounded Rationality, The American Economic Review, Vol. 81, No. 2, pp.353-359, 1991.

［Arthur95］ Arthur, W., Complexity in Economic and Financial Markets, Complexity, pp.20-25, 1995.

［EA96］ Epstein, J.M. and Axtell, R., Growing Artificial Societies: Social Science From The Bottom Up, The Brookings Institution, 1996.（邦訳：服部正太・木村香代子、人工社会　複雑系とマルチエージェントシミュレーション、株式会社構造計画研究所、1999）

［Bellman57］ R. Bellman, “Dynamic Programming,” Princeton University Press, Princeton, New Jersey, USA, 1957.

［CY96］ Chen, S.H. and Yeh, C.H., Genetic Programming and the Efficient Market Hypothesis, in Koza, J. et al. Eds., Genetic Programming 1996 : Proceedings of the 1st. Annual Conference, pp.45-53, the MIT Press, 1996.

［CK99］ Chen, S.H. and Kuo, T.W., Are Efficient Markets Really Efficient? : Can Financial Econometric Tests Convinve Machine Learning People?, in Proceedings of ICAI99, pp.444-450, CSREA, 1999.

［CT99］ Chen, S.H. and Tan, C.W., Brief Signals in the Real and Artificial Stock Markets: An Approach Based on the Complexity Function, in Proceedings of ICAI99, pp.423-429, CSREA, 1999.

［CY99］ Chen, S.H. and Yeh, C.H., On the Consequences of “Following the Herd”: Evidence from the Artificial Stock Market, in Proceedings of ICAI99, pp.388-394, CSREA, 1999.

［CYL99a］ Chen, S.H., Yeh, C.H. and Liao, C.C., Testing for Granger Causality in the Stock Price-Volume Relation: A perspective from the Agent-Based Model of Stock Markets, in Proceedings of ICAI99,

pp.374-380, CSREA, 1999.

[CYL99b] Chen, S.H., Yeh, C.H. and Liao, C.C., Testing the Rational Expectations Hypothesis with the Agent-Based Model of Stock Markets, in Proceedings of ICAI99, pp.381-387, CSREA, 1999.

[HM91] Holland, J. and Miller, J., Artificial Adaptive Agents in Economic Theory, American Economic Review, Papers and Proceedings, Vol. 81, No. 2, pp.365-370, 1991.

[Ho92] Holland, J., Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. 2 nd Ed., MIT Press., 1992.

[HT99] Holmstrom, B.R. and Tirole, J.（中村彰弘訳）、産業組織論ハンドブック第2章「企業の理論」②、郵政研究所月報、1999年4月.

[Koza92] Koza, J., Genetic Programming: On the Programming of Machines by Means of Natural Selection, MIT Press, 1992.

[Koza94] Koza, J., Genetic Programming II: Automatic Discovery of Reusable Programs, MIT Press.

[Minsky52] Minsky, M., A Neural-Analogue Calculator Based upon a Probability Model of Reinforcement, Harvard University Psychological Laboratories, Cambridge, Massachusetts, January 8, 1952.

[Samuel59] Samuel, A. L., Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, 3 (3), pp.211-229, 1959.

[Samuel67] Samuel, A. L., Some Studies in Machine Learning Using the Game of Checkers. II - Recent Progress, IBM Journal of Research and Development, 11 (6), pp.601-617, 1967.

[SB98] Sutton, R.S. and Barto, A.G., Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[UM00] U-Mart研究会HP、http://www.u-mart.econ.kyoto-u.ac.jp（2002年5月8日現在）.

[Watkins92] Watkins, C. and Dayan, P., Technical note: Q-Learning, Machine Learning 8 , pp.55-68, 1992.

[荒井他98] 荒井幸代・宮崎和光・小林重信、マルチエージェント強化学習の方法論：*Q*-Learningと Profit Sharingによる接近、人工知能学会誌、Vol. 13、No. 5、pp.609-618、1998年9月.

[荒井01] 荒井幸代、マルチエージェント強化学習：実用化に向けての課題・理論・諸技術との融合、人工知能学会誌、Vol. 16、No. 4、pp.476-481、2001年7月.

[和泉他00] 和泉潔・植田一博、人工市場入門、人工知能学会誌、Vol. 16、No. 6、pp941-950、2000.

[川村他00] 川村秀憲・車谷浩一・大内東、X-Economy －マルチエージェント経済におけるシミュレーションプラットホームー、第10回マルチ・エージェントと協調計算ワークショップ（MACC2001）論文集、pp.122-127、2001.11.16-17.

[木村他99] 木村元・宮崎和光・小林重信、強化学習システムの設計指針、計測と制御、Vol. 38、No. 10、1999.

[塩沢00a] 塩沢由典、V-Martの意義　共通テストベッドとしてのバーチャル市場、http://www.u-mart.econ.kyoto-u.ac.jp/articles/u-mart-importance.pdf（2002年5月8日現在）.

[塩沢00b] 塩沢由典、経済学にとっての人工市場、人工知能学会誌、Vol. 16、No. 6、pp.951-957、2000.

[谷田01] 谷田則幸、分子生物学とコンピュータサイエンスの接点、関西大学情報処理センターフォーラム、No. 17、関西大学情報処理センター、2002年3月.

[西村98] 西村和雄、複雑系経済学とは何か、東京情報大学研究論集、Vol. 2、No. 3、pp.147-168、1998年4月.

[保知他02] 保知良暢・松井藤五郎・犬塚信博・世木博久、マルチエージェント強化学習における報酬発生条件に基づく貢献度判別と報酬分配、2002年度人工知能学会全国大会(第16回)、2D3-02、2002年5月.

[宮崎他94] 宮崎和光・山村雅幸・小林重信、強化学習における情報割当の理論的考察、人工知能学会誌、Vol. 9、No. 4、pp.104-111、1994.

[宮崎他99a] 宮崎和光・荒井幸代・小林重信、POMDPs環境下での決定的政策の学習、人工知能学会誌、Vol. 14、No. 1、pp.148-156、1999年1月.

[宮崎他99b] 宮崎和光・荒井幸代・小林重信、Profit Sharingを用いたマルチエージェント強化学習における報酬配分の理論的考察、人工知能学会誌、Vol. 14、No. 6、pp.1156-1164、1999年11月.