間接的互恵関係モデルにおける価値観の進化

岩　島　奈緒美[*]　・　有　田　隆　也[**]

Evolution of sense of value in a model of indirect reciprocity

Naomi IWASHIMA[*]　and　Takaya ARITA[**]

Abstract

Darwinian evolution should provide an explanation for cooperative behavior. Nowak and Sigmund present a theoretical framework that is based on indirect reciprocity, and conclude that cooperation have evolved through indirect reciprocity by image scoring. We focus on the evolution of sense of value by introducing diversity in evaluation of altruistic behavior of other individuals into their model from the viewpoint of evolutionary psychology. This paper investigates its effect on emergence and maintenance of cooperative relationship in a population of individuals based on the results of simulation experiments.

Key word: 間接的互恵関係，価値観の多様性，協調の進化，イメージスコアリング，進化心理学

1 はじめに

自己複製子としての遺伝子は，長生きで，多産性で，複製が正確であるという３条件を満たす，言い換えるならば「利己的」な性質を備えるものが結果的に集団に広がるとみなすことができる[Dawkin 1989]．一方，その遺伝子の「乗り物」である個体レベルにおいては，道徳的な人間社会だけでなく，動物の群れの間にも利他的行為を観察することができる．自己の遺伝子の保存のために利己的に振舞いがちと考えられる個体が，なぜ利他的行為を見せるのだろうか？　さらに，利他性がその基盤を形作っていると考えられる我々の社会は，人間という種が進化していく過程において，どのような条件で創発し，また維持され得てきたのだろうか？　このような問いかけは，生物学，社会学の領域から心理学，ゲーム理論，計算

[*] 名古屋大学大学院人間情報学研究科（博士前期課程）
Graduate School of Human Informatics, Nagoya University (Master's Course)
[**] 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University

機科学の領域に至る広範囲において議論されてきたテーマであり，長年に渡って様々な研究・議論がなされてきた[Riolo，Cohen，Axelrod 2001]．

言語能力を備えた人間は，文化の力によって遺伝子の影響から逃れ，利他的行為を駆動しているという側面は無視することができない．倫理，道徳，教育，法律などである．しかし，そのような高いレベルの概念を使わずに生物学レベル，あるいは数学レベルで人間に留まらない利他的行為を説明することがもし可能ならば，それは利他性をより普遍的に理解できたことを意味する．そのような観点からの生物学的な説明として，主に「血縁選択」「グループ選択」「互恵関係」という３つのメカニズムが提案されてきた．

まず，「血縁選択」による説明とは，血縁者相手に利他的行為を行うと，その個体にとっては損失であるが，共通の遺伝子を増やす（包括適応度を高める）ことができるため，個体レベルでは利他的であっても，遺伝子の観点からはその「利己性」を保てるというものである[Hamilton 1964]．次に，「グループ選択」に基づく説明とは，あるグループに属する個体が利他的行為を行うと，その個体にとっては損失であるが，そのグループ全体の適応度が上がることで，複数のグループがグループ単位で競っているという状況下では，グループ選択において淘汰されにくくなるため，というものである[Williams 1971]．一方，「互恵関係」に基づく説明とは，利他的行為を行った個体は，見返りを考慮するとその行為による犠牲以上の利益を得られるため，というものである．

「互恵関係」に基づく説明は，利他的行為を行う側と受ける側で特別な関係を必要としない点が特長であり，さかんに研究が行われてきた．その中でも，先駆的研究が Axelrod による繰り返し囚人のジレンマゲーム(IPD)を用いた協調戦略の研究である[Axelrod 1984] [有田 2002]．彼は，ある二者間でゲーム論的な状況設定において相互作用が繰り返される場合，互恵関係に基づいて利他的行為（「しっぺ返し戦略」）が集団中に広まりうることを示した．互恵関係に基づく利他的行為は，我々の生活において自然なものと考えられるだけでなく，さまざまな生物種における生態系においても実在するとの報告がなされている（たとえば[Milinski 1987]）．しかし，一方で，そのような説明が基盤とする囚人のジレンマゲームが１対１という設定であり，助けた（協調した）相手から直接的な見返りを期待する直接的互恵関係であるので，説明する範囲が同じ二者間で繰り返し相互作用を行う場合に限定される．つまり，大規模な集団における利他的行為，端的に言えば，今後二度と会わないような相手に対する利他的行為については説明できない．

この限界を超えるために，近年，Nowak らによって，イメージスコアに基づく間接的互恵関係に関するモデルが提案された[Nowak，Sigmund 1998]．これは，同じ二者間での相互作用の繰り返しを必要としないモデルである．彼らはこのモデルに関する進化シミュレーションによって，協調的集団と裏切り集団が周期的に出現するという基本的な進化ダイナミクスを観察し， Axelrod らの説明し得なかった，比較的大規模な集団における利他的行為の基盤

を説明することに成功している．彼らのモデルにおける協調成立の要因は，一言で言うならば，相手を助ければ，その行為によって受援者に留まらない他者において自分のイメージ（評判）が高まり，将来，自分が受援者になる可能性が高まることが期待できるということである．利他的行為を行った相手からの直接的な恩返しだけが期待できるのではなく，その行為を知った人からの利他的行為も期待できるという点に意義がある．

Nowak らのこの間接的互恵関係に関するモデルはその後，様々な観点から引き続き研究が行われている．北野らは，このモデルに対して Dunbar のゴシップ説的な意味合いのコミュニケーション[Dunbar 1996]を導入している[北野，有田 2000a,b]．Nowak らのモデルに対し，「誰々が利他的行為を行った」という情報交換ができるように拡張したものである．その際，実際はしていない利他的行為を自分がしたと欺きを行えるようにして，嘘つきの進化も検討の対象とした．実験の結果，1）グループサイズが増加していくと，嘘のようなコミュニケーションのネガティブな面が働いて互恵関係維持が難しくなるが，2)その後，次第に嘘つきが駆逐されていくこと,3）嘘つきが蔓延している状態でも互恵関係が一定の割合で成立していることなどが示された．また，Lotem らは，現実世界を見ると，ハンディキャップや病気である，あるいは若すぎるなどの理由で，遺伝子型ではそうでなくても，表現型として利他的行為を行えないという人が一定の割合で存在するという考えから,「表現型裏切り者」を導入して実験した．その結果，逆説的であるが，識別者の割合が増加し，Nowak らの実験で見られたような協調的集団と裏切り集団のサイクルはなくなり，識別的集団によって，むしろ安定した協調関係が築かれることが示された[Lotem，Fishman，Stone 1999]［Badcock 2000］．他にも，人間を被験者とした心理実験によって，このような間接的互恵関係が実在することを検討した研究もなされている[Wedekind，Milinski 2000]．

本論文は，そのような一連の拡張研究とは異なり，Nowak らによるイメージスコアに基づく間接的互恵関係モデルの基本部分の妥当性，特に各個体の心理に関する進化的妥当性に関して，シミュレーション実験に基づいて計算論的アプローチを行うことを目的とする．Nowak らのモデルでは，利他的行為（支援）は，受援者に対するイメージを表す値（各個体が他個体ごとに持つ）が，支援するかどうかを判断するための閾値を表す値（各個体が一つだけ持つ）以上の場合に行われる．この閾値が各個体によって異なる設定となっているために，Nowak のモデルにおいても，この値の大小によって，支援しがちな個体とそうでない個体という意味での多様性が進化的に生じる設定となっていた．しかし，そもそも，利他的行為を目撃したときに，一律にその支援者のイメージが１だけ増加するということがアプリオリに規定されていた．この部分に個体の持つ価値観の多様性を導入し，その価値観がどのように進化するかということを検討することを目指そうというのが本研究の基本的なモチベーションである．たとえば，利他的行為を目撃した場合，利他的行為者のイメージは全目撃者において一律に１だけ増加するのではなく，その増加の度合いは人様々であろうし，逆にイ

メージを落とすことも場合によってはありうるかもしれないということである．このような面に関して，心の機能の先天性と適応性に基づく進化心理学的なスタンスから価値観の多様性に焦点を合わせる．そして，各個体における価値観がどのように進化し，またその進化が協調的社会の創発や維持にどのような影響を与えるかを個体間の相互作用に基づくエージェントベースモデリング手法に基づく進化シミュレーション実験により明らかにする．なお，本論文では，先行研究と同様に，「利他的行為」とは，動機の心理に関わるものではなく，当の行為の結果として，利他的行為者の（生物学的な意味における）適応度を低め，同時に受益者の適応度を高める行為と定義する．

2　モデル

2 - 1　Nowak らのモデル

集団は n 個体から構成され，2 個体間で支援-受援の機会が繰り返し起こる．各個体は，支援するかどうかの閾値である戦略値 k と，集団内の他個体に対するイメージスコア s を他個体ごとに持つ．また，適応度に相当する得点 q も持つ．

まず，集団から 2 個体が，それぞれ支援予定者・受援予定者としてランダムに選ばれる．支援予定者は，もし，

受援予定者のイメージスコア　≧　支援予定者の戦略値

であるならば支援する．この際，

・支援者はコスト c が得点からマイナスされ，受援者は利益 b（$b>c$）が得点にプラスされる．

・（受援者を含む）各目撃者において，支援者に対するイメージスコアが 1 上がる．

もし，

受援予定者のイメージスコア　＜　支援予定者の戦略値

であるならば支援しない．この際，

・得点の増減はない

・（受援者を含む）各目撃者において，支援者に対するイメージスコアが 1 下がる．

つまり，これは，支援（利他的行為）を行うとコストを払わなければならないが，それによって皆の自分に対するイメージが上がれば，将来自分が支援されやすくなり，コスト以上の利益を得られる可能性がある，という構造をモデル化したものである．

上記のように二個体ずつランダムに選んで相互作用を行うことを，一世代につき m 回繰り返す．世代の終わりに，個体の総数は固定したまま，各個体は得点に比例した数だけ自分の戦略値 k を受け継ぐ子孫を残し，イメージスコアと得点はリセットする．但し，戦略値 k に対して，突然変異（1/1000 の確率で子はランダムな戦略値を取る）を導入する．以上の処理を 1 世代として，これを t 世代繰り返す．

Nowak らは，パラメータを n=100，b=10，c=1，k=-5，-4，・・，5，6，m=300，t=100000 と設定して，実験を行った．その結果と分析結果の概要は次の通りである．

・全世代を通した平均協調率（支援を行った割合）は 70 パーセント程度であり，過去にほとんど面識のない二者間でも利他的行為が行われ得ることが示された．

・ほとんどの世代で，識別戦略（戦略値 k=0,-1）が多数を占める社会になるが，極端な協調戦略（戦略値 k=-5,-4）が増加し，そのために極端な裏切り戦略（k=5,6）が侵入可能になる．しかし，裏切り戦略が蔓延すると，すぐにまた識別戦略が多数を占める社会に戻ることを繰り返し，協調と裏切りの無限のサイクルを見せる．これは，識別戦略が多数を占める社会では，極端な裏切り戦略の侵入は不可能であるが，突然変異の影響で極端な協調戦略が増加すると，そのような社会ではただのりに制裁を加える機能が無いために，極端な裏切り戦略の侵入を許してしまうからである．これと同様な協調と裏切りのサイクルは，繰り返し囚人のジレンマゲームに関する戦略の進化実験（たとえば[Lindgren 1992]）においても見られる現象である．

2-2 モデルの基本部分の変更

Nowak らのモデルでは，支援するのを目撃すると目撃者の支援者に対するイメージスコアは 1 上がるとアプリオリに設定されていた．本モデルでは，その値を遺伝的なパラメータ（イメージスコア操作基準値 std と呼ぶ）として進化させることを可能とする．よって，各個体は集団内の他者それぞれに関するイメージスコア s，戦略値 k，得点 q の他に，イメージスコア操作基準値 std を持つことになる．相互作用の際，支援するのを目撃すると，目撃者の支援者に対するイメージスコアが 1 上がるのではなく，各目撃者が固有に持つ値 std ずつ上がり，支援しないのを目撃すると，1 下がるのではなく，各目撃者が固有に持つ値 std ずつ下がるという設定に変更する（図 1）．これにより，std がマイナスの個体の場合には，その支援を偽善的行為とみなしたかのごとく，行為者に対するイメージが逆に下がるわけである．そして，子孫を残す際には，自分の戦略値 k だけでなく，イメージスコア操作基準値 std も遺伝させる．

これにより，Nowak らのモデルでは，戦略値 k だけを個体ごとに持たせて，他者を支援するかどうかの判断に多様性を持たせていたが，本モデルでは，それだけではなく，イメージスコア操作基準値 std を導入することにより，他個体の支援を行った，あるいは行わなかったという事実を評価する際にも，人間の価値観の多様性を反映させることを可能とした（図 2，図 3）．そして，それらがどのように進化し，また，集団としてどのような振舞いを見せるのかを観察することにした．

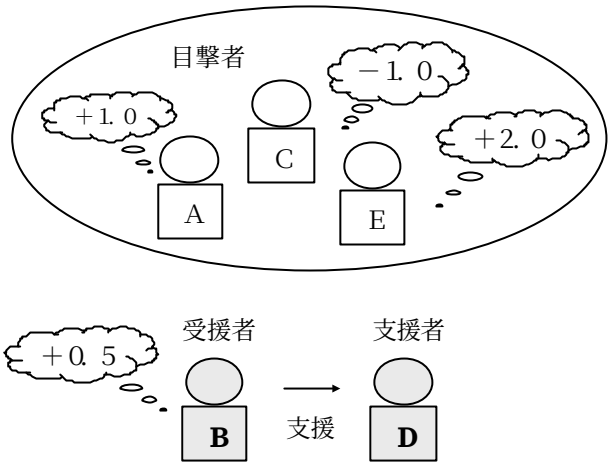

図１　価値の多様性のイメージ

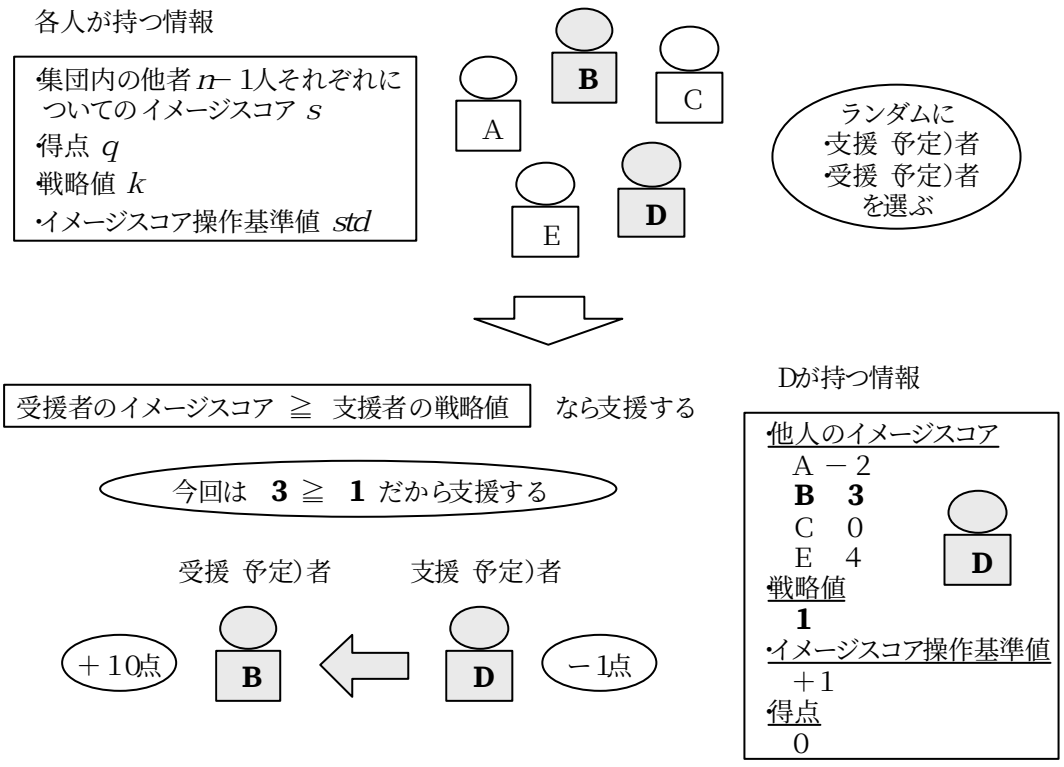

図２　各個体の持つ情報と支援条件

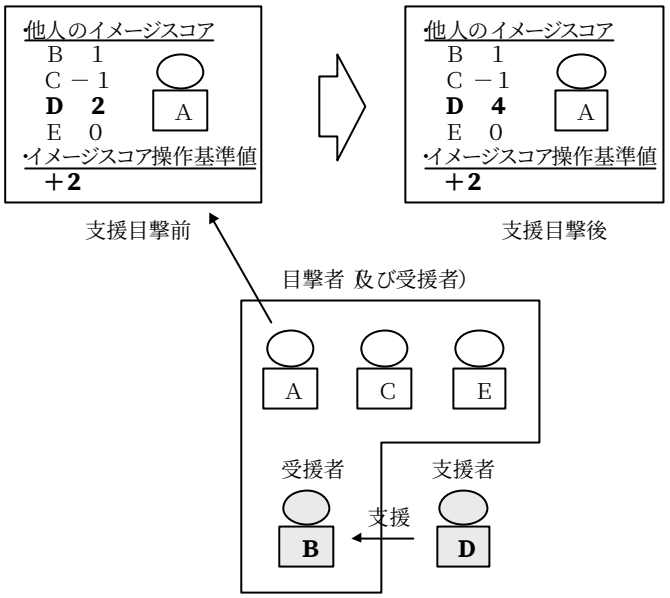

図３　支援目撃前後のイメージスコア変化

３　シミュレーション実験

3-1実験Ａ　全個体が同じ *std* 値を持つ場合

まず，全体的な傾向を調べるために，イメージスコア操作基準値 *std*（支援を目撃した場合のイメージの増減）について，個別に与えるのではなく，全員が同じ *std*（= *w_std*）を持つ設定にして実験を行った．パラメータは，*n*=100，*b*=10，*c*=1，*k*=-5～6 で 1.0 刻み（最初の世代ではランダムに与える），*m*=300，*t*=100000 を用いた．

図 2-1 から図 2-7 は，*w_std* を-1.0 から 2.0 まで 0.5 刻みで変化させた場合の戦略値の推移を示している．同図より，*w_std*≦0 だと平均戦略値が負にならない（協調戦略があまり出現しない）ことや，*w_std*＞0 なら平均戦略値が負になる世代が見られるが，*w_std* が大きくなっても上記のような協調と裏切りの無限サイクルは無くならないことがわかる．

また，表 1 は，各場合における協調率（支援が行われた割合）を示している．同表より，*w_std*=1.0 近辺を境にして，それ以上ならば協調的社会（協調率 50%以上），それ以下（マイナスも含む）なら裏切り社会（協調率 50%以下）が形成されることがわかる．*w_std*=1.0 で分かれるのは，戦略の刻み幅に依存している部分もあり，1.0 未満だと，一度の協調（又は裏切り）がすぐには後の支援（又は不支援）につながらないためであると考えられる．また相互作用回数 *m* もこの結果に深く関わっていると推測される．

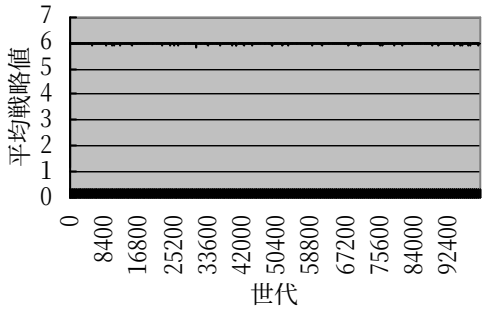

図4-1 平均戦略値の推移
(*w_std*=-1.0)

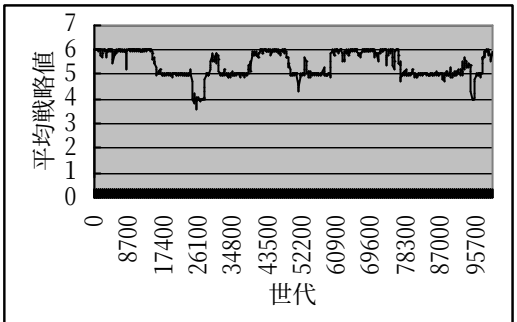

図4-2 平均戦略値の推移
(*w_std*=-0.5)

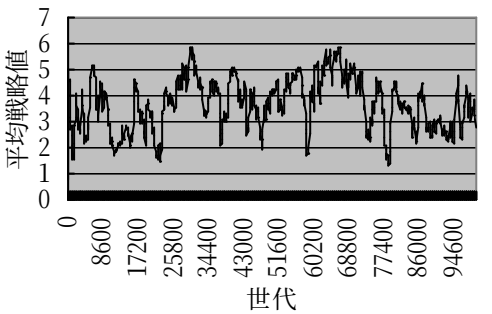

図4-3 平均戦略値の推移
(*w_std*=0.0)

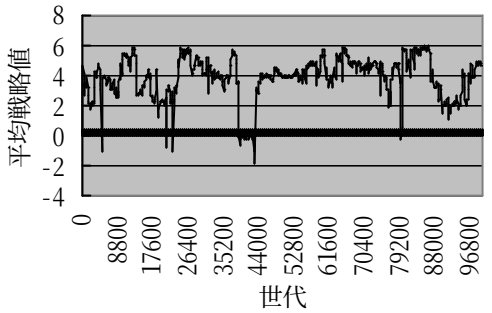

図4-4 平均戦略値の推移
(*w_std*=0.5)

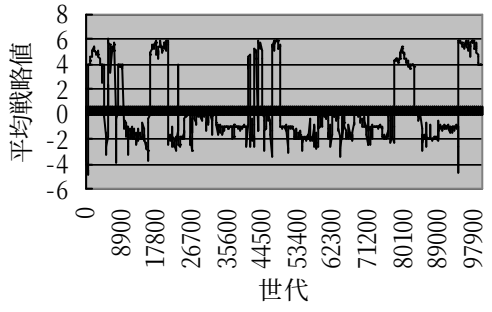

図4-5 平均戦略値の推移
(*w_std*=1.0)

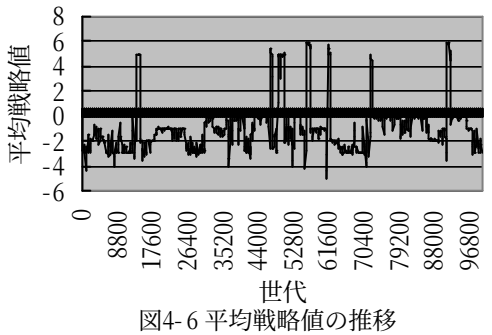

図4-6 平均戦略値の推移
(*w_std*=1.5)

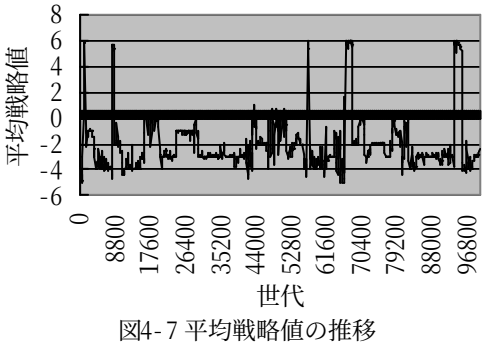

図4-7 平均戦略値の推移
(*w_std*=2.0)

表 1　w_std と協調率

w_std 値	-1	-0.5	0	0.5	1	1.5	2
協調率（%）	0.0763	0.0812	0.0565	4.6967	70.688	92.265	91.691

そこで，次に，世代あたりの相互作用回数 m と全体統一のイメージスコア操作基準値 w_std の両者を同時に変えて，全世代を通した協調率を調べた．図 3 は，世代あたりの相互作用回数 m と全体統一のイメージスコア操作基準値 w_std を変化させた場合の，全世代を通した協調率を表している．但し，全世代の合計相互作用回数を 3000 万回に統一し，パラメータは，n=100，b=10，c=1，k=-5～6 を用いた．

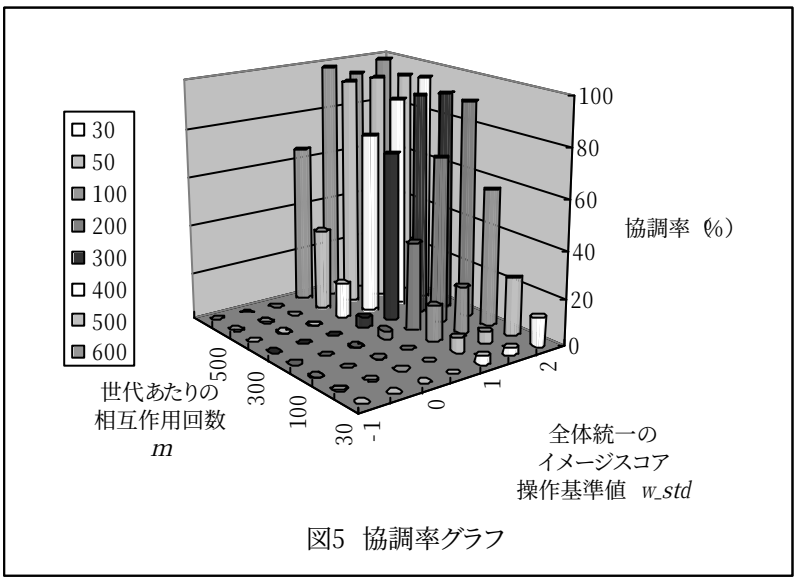

図5　協調率グラフ

その結果，　$w\_std \leqq 0$ の場合は，相互作用回数を増やしても支援はほとんど行われなかった．これは，支援すればするほどイメージスコアが下がり，将来支援してもらいにくくなるためであると考えられる．そして，$w\_std > 0$ の場合は，w_std の値が大きくなるにつれ，少ない相互作用回数でも協調的社会（支援率 50 パーセント以上の社会）が築かれるようになった．これは，一度の支援の効果が大きくなるためであろう．

これらの結果より，利他的行為に対する各個人の持つ心理的な評価が低いと，利他的行為は行われにくく，逆に，利他的行為に対する各個人の持つ心理的な評価が高まれば，少ない相互作用回数でも協調社会が築かれるため，より大規模集団においても協調社会が創発しうるということが示された．

3-2実験Ｂ　std 値を進化させた場合

本実験では，イメージスコア操作基準値 *std* について，個別のイメージスコア操作基準値（*i_std*）を与え，更に，その *i_std* の突然変異（1/1000 の確率で，子はランダムな *i_std* 値を取る）を導入した．但し，パラメータは，n=100，b=10，c=1，k=-5～6，m=300，t=100000 を用いた．

図４に，*i_std* の取りうる範囲を変化させた場合の協調率を示す．同図より，*i_std* のとりうる範囲の上限が 0 以下ならば協調的社会は不成立であり，下限が 1 以上ならば協調的社会が築かれることがわかる．そして，様々な *i_std* が混在するような社会では，*i_std* の取りうる範囲の平均が 1.0 以上ならば，協調的社会（協調率 50％以上）が形成された．それ以下だと，協調的社会にはなりにくくなる．これは，相互作用回数mを 300 回に固定していることにも依存すると考えられるが，いずれにせよ，0 以下の *i_std* を持つ個体が存在しても協調的社会が創発しうるという事実が示された．

これらの結果より，利他的行為に対する評価基準に個人の価値観の多様性を取り入れた場合でも協調的社会が創発しうるということが示された. 更に, 創発した協調的社会の中には, 利他的行為に対して良い評価をしないような個体も存在するような多様な社会が存在していることは注目すべき点であろう.

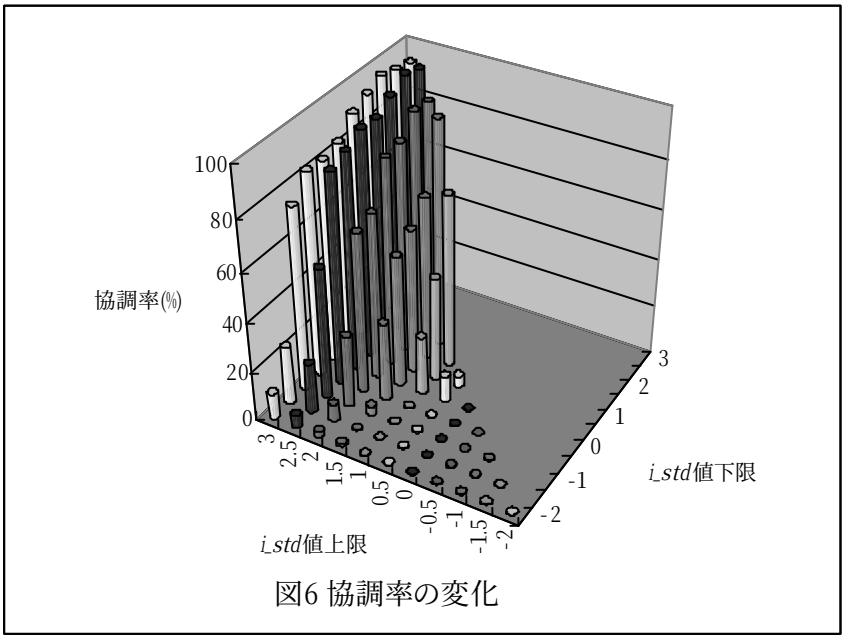

図6 協調率の変化

3-3価値観の多様性が社会に与える効果の分析

本実験では，価値観の多様性が社会に与える効果について検証する．具体的には，実験Ｂで定義した *i_std* の取りうる値の範囲の平均を 1.0 に保ったまま多様性を変更した場合の影

響を調べた.

図５-1 から図５-5 は，実験Ｂにおける，①*std*=1.0，②0.5≦*i_std*≦1.5，③0.0≦*i_std*≦2.0，④-0.5≦*i_std*≦2.5，⑤-1.0≦*i_std*≦3.0 のそれぞれの場合における *i_std* 値ごとの戦略値の出現頻度を全世代で平均した結果をグラフである．同図において，丸の半径を出現頻度に比例させて表している.

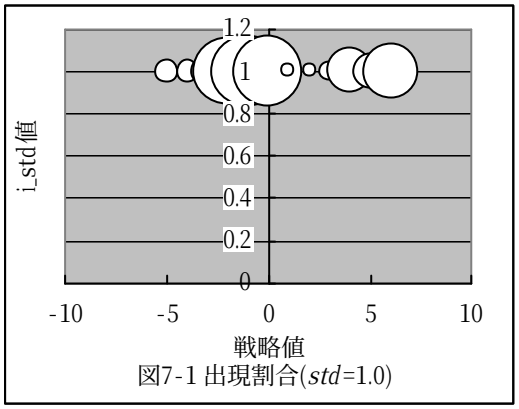

図7-1 出現割合(*std*=1.0)

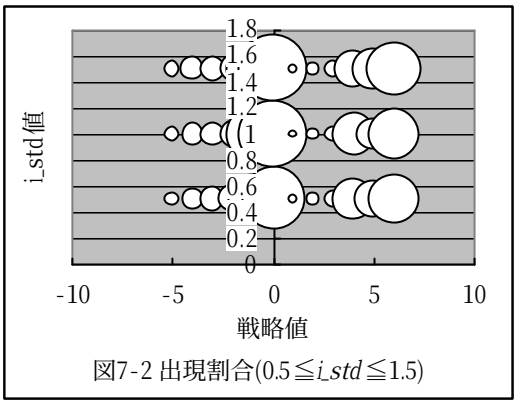

図7-2 出現割合(0.5≦*i_std*≦1.5)

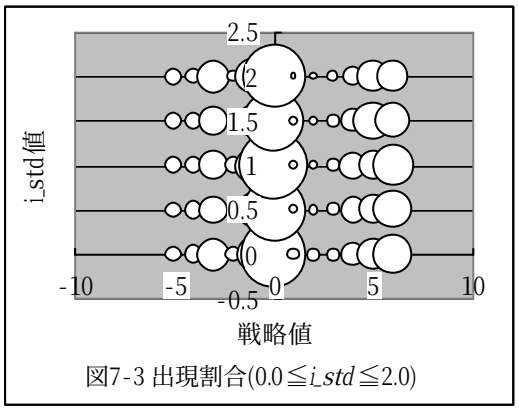

図7-3 出現割合(0.0≦*i_std*≦2.0)

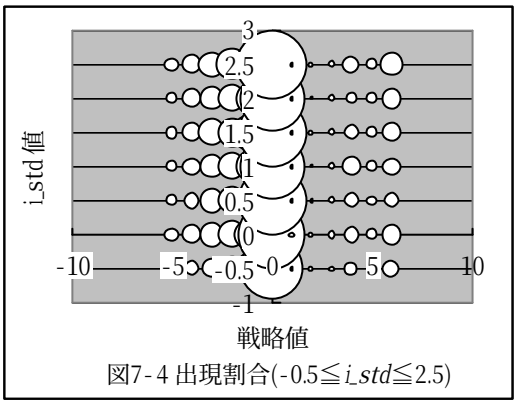

図7-4 出現割合(-0.5≦*i_std*≦2.5)

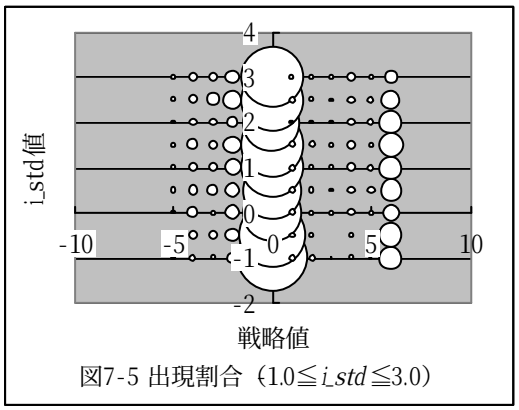

図7-5 出現割合（1.0≦*i_std*≦3.0）

同図より，どの場合も全世代を通してみれば、基本的には識別戦略（戦略値 k=0,-1）の出現割合が最も高く，次いで極端な協調戦略（戦略値 k=-5,-4）が多く出現し，極端な裏切り戦略（k=5,6）もいくらか出現しているのがわかる．しかし，多様性が大きくなる（i_std の取りうる値の範囲が広くなる）ほど，完全な識別戦略（k=0）の出現割合が高くなっている．これは，利他的行為に対して良い評価をしないような者は，協調戦略をあまり助けず，逆に利他的行為に対してかなり良い評価をするような者は，裏切り戦略をあまり助けないので，識別戦略が有利になるのだと考えられる．つまり，利他的行為に対する評価基準の多様性が不正を抑制する，つまり，ただ乗りを許さないような頑健性を生むということが言える．

4　おわりに

本論文では，Nowak らによるイメージスコアに基づく間接的互恵関係モデルの基本部分の妥当性，特に各個体の心理に関する進化的妥当性に関して，シミュレーション実験に基づいて分析した．イメージスコアによる間接的互恵関係モデルの特徴として，同一個体との直接的な相互作用がほとんど無くても，協調関係が創発するという点があげられるが，協調的社会が築かれるための一世代あたりの相互作用回数の最低値は，全体統一のイメージスコア操作基準値に依存し，さらに，戦略の刻み幅による効果の表れやすさに，大きな影響を受けていることが示された．効果が現れやすいと，相互作用がより小さくても協調関係が成り立つということは，より大規模な協調集団（社会）を構成しうるということである．したがって，これらの知見は，人々のつながり，つまり社会性が高まるためには，利他的行為に対する各個人の持つ心理的な評価が高まることが必要であるということを示唆している．

また，個別に様々なイメージスコア操作基準値を与えて進化させることによって，利他的行為に対する評価基準に個人の価値観の多様性を導入した場合においても，協調的社会が創発しうるということがわかった．創発した社会の中には，利他的行為に対して良い評価をしないような個体も存在するような多様な社会も観察された．そして，利他的行為に対する評価基準が，個人差はあっても平均すれば Nowak の設定と同じであるような社会において，基本的に協調社会となり，協調と裏切りのサイクルも見られたことから，Nowak らの元の設定は，協調的社会の創発・維持という面から見れば妥当であることが示された．さらに，社会の構造という観点からは，利他的行為に対する評価基準の多様性が増すほど識別戦略が多く出現する点は興味深い．これは，他者イメージを形成する心理的機能の多様性が増すと，識別レベルが上がり，協調的社会が維持されるためである．他者イメージを形成する心理的機能の多様性が，社会にただ乗り戦略を許さないような頑健性を生み出すということは特筆すべき点である．

参考文献

[1] 有田隆也，*人工生命（改訂２版)*，医学出版（2002）.

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

[3] C. Badcock, *Evolutionary Psychology,* polity (2000).

[4] R. Dawkins, *The Selfish Gene*, Oxford University Press (1989).

[5] R. Dunbar, *Grooming, Gossip, and the Evolution of Language*, Faber & Fabe (1996).

[6] W. D. Hamilton, "The Genetic Evolution of Social Behavior", *Journal of Theoretical Biology*, pp. 1-52 (1964).

[7] 北野純，有田隆也，“情報交換が可能な間接的互恵関係の進化モデル”，*「人工生命新しい潮流」研究会論文集*（計測自動制御学会第 20 回システム工学部会研究会資料）pp．99-104（2000a）.

[8] 北野 純，有田隆也,"情報交換が可能な間接的互恵主義の進化シミュレーション"，第 29 回数理社会学会大会研究報告要旨集，pp．12-15（2000b）.

[9] K. Lindgren, "Evolutionary Phenomena in Simple Dynamics", *Artificial Life II*, pp.295-312 (1992).

[10] A. Lotem, M. A. Fishman and L. Stone, "Evolution of cooperation between individuals", *Nature*, 400, pp.226-227 (1999).

[11] M. Milinski, "Tit for Tat and the Evolution of Cooperation in Sticklebacks", *Nature*, 325, pp.433-435 (1987).

[12] M. A. Nowak and K. Sigmund, "Evolution of indirect reciprocity by image scoring", *Nature*, 393, pp.573-577 (1998a).

[13] M. A. Nowak and K. Sigmund, "The Dynamics of indirect reciprocity", *Journal of Theoretical Biology*, 194, pp.561-574 (1998b).

[14] R. L. Riolo, Mi. D. Cohen and R. Axelrod, "Evolution of cooperation without reciprocity", *Nature*, 414, pp.441-443 (2001).

[15] C. Wedekind and M. Milinski, "Cooperation Through Image Scoring in Humans", *Science*, 288, pp.850-852 (2000).

[16] G. C. Williams, *Group Selection*, Aldine-Atherton, Chicago (1971).