マルチエージェントシステムにおける利他的な行動規則の獲得

上田　祐彰[†]　谷澤　俊彰[††]　高橋　健一[†]　宮原　哲浩[†]

Acquisition of Reciprocal Altruism in a Multi-Agent System

Hiroaki UEDA[†], Toshiaki TANIZAWA[††], Kenichi TAKAHASHI[†], and Tetsuhiro MIYAHARA[†]

あらまし　マルチエージェントシステムにおける利他的な行動規則の獲得手法について論じる．はじめにタスク割当問題を定義し，この問題において公平にタスクを分担するような行動規則を獲得するためには，利他的な行動規則の学習が重要であることを示す．次にタスク割当問題のマルチエージェントモデルを提示し，Q 学習を用いた行動規則の獲得手法を提示する．本論文では，利他的な行動を獲得する要因として，他のエージェントの不幸を悲しむ感情である憐憫悲，及び他のエージェントの幸福を喜ぶ感情である憐憫喜を導入し，これらが利他的な行動規則の獲得に及ぼす影響について考察する．最後に計算機シミュレーションの結果を提示し，適切な強さの憐憫悲の導入は利他的な行動規則の獲得に寄与するが，憐憫喜及び過度の憐憫悲の導入は利他的な行動規則の獲得には寄与しないことを示す．

キーワード　マルチエージェントシステム，強化学習，利他的な行動

1. まえがき

サッカーゲームや国際商取引などの社会活動をマルチエージェントシステムによってモデル化し [1]〜[3]，強化学習 [4]〜[6] や進化論的計算手法 [7]〜[9] を用いてエージェントの行動規則を獲得するための研究が多くの研究者によってなされている．マルチエージェントシステムにおける行動規則の学習技法の多くは，協調行動の獲得による報酬の最大化を目的としているが，これらの技法で獲得される規則は利己的，すなわち各エージェントの得る報酬の最大化を達成するためのものである場合が多い．一方現実の人間社会においては，献血やボランティア活動などに代表されるような，個人の利益にはならないが他人あるいは社会全体に利益をもたらす利他的な行動が，文化的な人間社会の形成にとって重要な役割を担っている．利他的な行動を選択するための要因には，利他的な行動が社会全体の利益となることを知っていること，自己満足あるいは同情などの感情，他人の自分に対する評価など，様々な要因が考えられる．本論文では，同情や嫉妬などの感情に着目し，計算機実験を通じて感情が利他的な行動の選択に及ぼす影響について考察する．

はじめに，人間社会の簡単なモデルとしてタスク割当問題 [10] を定義する．ここで定義するタスク割当問題は，複数のメンバからなるコミュニティに対して数種類のタスクが逐次的に提示される問題であり，すべてのタスクを公平かつ効率的に処理することを目的としている．各メンバは，現在提示されているタスクの種類と他のメンバの状態に基づいて，そのタスクを処理するか否かを決定し，あるメンバがタスクを処理するとそのメンバに報酬が与えられる．タスクには，処理の完了に多くの時間を必要とするハードタスクと，容易に完了できるタスクがあり，ハードタスクを多く処理したメンバは獲得する総報酬が少なくなるように定義されている．このため，すべてのメンバが利己的な行動を選択する場合には，タスクの公平な分担及び報酬の公平な分配が困難となる．したがってタスク割当問題の目的を達成するためには，自分自身の報酬は減少するが他のメンバのためにハードタスクを処理する行動，すなわち利他的な行動を選択可能な行動規則（利他的な行動規則）の獲得が必要である．

次に，タスク割当問題のマルチエージェントシステムによるモデルを提示し，強化学習による行動規則の

[†] 広島市立大学情報科学部知能情報システム工学科，広島市
Faculty of Information Sciences, Hiroshima City University, Hiroshima-shi, 731–3194 Japan
[††] 京都大学大学院情報学研究科，京都市
Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto-shi, 606–8501 Japan

学習手法について論じる．ここでは，他のメンバの幸福を喜ぶ感情，及び他のメンバの不幸を悲しむ感情の 2 種類の感情を導入し，感情による影響を報酬に反映させることにより，利他的な行動規則の獲得を試みる．また，本論文で提示するマルチエージェントモデルを計算機上に実装し，シミュレーション実験を行うことにより，感情が利他的な行動規則の獲得に及ぼす影響について考察する．

本論文は次のように構成されている．**2.** では，本論文と関連する研究について概観する．**3.** では，本論文で対象とするタスク割当問題を定義する．**4.** では，タスク割当問題のマルチエージェントシステムによるモデルを示し，強化学習による行動規則の学習手法について述べる．**5.** では実験結果を提示し，感情が利他的な行動規則の獲得に及ぼす影響について検討する．

2. 関 連 研 究

マルチエージェント環境における強化学習に関する研究の多くは，各エージェントあるいはエージェント社会全体の利益を最大化することを目標として，個々のエージェントが自律的に行動規則を学習する状況を想定している．また近年では，エージェント間での分業や利他的な行動の学習により，エージェント社会全体の利益を最大化する手法についても研究されている．

各エージェントの受ける報酬の調整は，エージェント間での分業あるいは利他的な行動を獲得する上で重要な役割を担っている．西岡らはこれに着目し，モチベーションパラメータを用いた報酬調整によってエージェント間での分業を実現しており，各エージェントの満足度や社会全体の利益を向上させている [11]．森山らは，エージェントのおかれている状況を非干渉状況，泥沼状況，競合状況の 3 種に分類し，状況に応じてエージェントの報酬を自動的に調整する手法を提案している [12]．彼らの分類では，すべてのエージェントが利己的であるときに社会全体が最良となる状況を非干渉状況，すべてのエージェントが利他的であるときに社会全体が最良となる状況を泥沼状況，一部のエージェントが利己的かつ残りが利他的であるときに社会全体が最良となる状況を競合状況と定義している．森山らの手法は，競合状況のモデルである狭路問題に対して特に良好な結果が得られることが報告されている．

筆者らは，森山らの提唱している競合状況は三つの状況に細分類されると考えている．一つ目は，エージェント群が利己的なものと利他的なものに二極化されるとき，すなわちエージェント間での一種の分業が獲得されたときのみ社会全体が最良となる状況である．本論文では，この状況を分業的競合状況と呼ぶ．二つ目は，各エージェントが利己的な行動と利他的な行動をバランス良く選択したとき，すなわち，すべてのエージェントが互恵的であるときのみ社会全体が最良となる状況である．本論文では，この状況を互恵的競合状況と呼ぶ．最後の分類は，分業的競合状況，互恵的競合状況のいずれにも属さない競合状況であり，本論文ではこのような状況をその他の競合状況と呼ぶ．すなわちその他の競合状況とは，エージェント間での一種の分業が獲得されるか，すべてのエージェントが互恵的であるときに，社会全体が最良となる状況である．このような分類を用いると，森山らの使用した狭路問題は分業的競合状況に分類される．その他の競合状況に分類される代表的な問題には，追跡問題 [13] がある．渡辺らは，互恵的な概念に基づいて利他的な行動を学習する手法を提案しており，追跡問題に対して良好な結果を得ている [13]．しかし [13] では，エージェント間での分業が獲得されたことによる結果であるのか，互恵的なエージェントが学習されたことによる結果であるのかが明示されていない．そこで本論文では，互恵的競合状況に分類される問題としてタスク割当問題を定義し，エージェントによる互恵的な行動の学習により，エージェント社会全体に対する利益の最大化を試みる．

複数のエージェントによる協調行動の実現や，利他的な行動を学習する上での代表的な動機付け方法としては，交渉などのエージェント間の相互作用が挙げられる．交渉はエージェント間の合意形成にとって重要な役割を果すが，囚人のジレンマ [6] や狭路問題 [12] のように，これを導入すべきではない問題も存在する．更に，感情をもった人間によって構成されている現実社会，特に日本的な社会では，自発的に協力することを美徳とする，協力の要請を恥と考えるなど，交渉による利他的な行動の動機付けが有効とはならない場合もある．したがって本論文では，利他的な行動を選択するための動機付けは，同情などの感情を用いた報酬調整により行っている．また，モチベーションパラメータ [11] のように協調行動や利他的な行動に対する動機付けを明示的に与えるのではなく，感情に基づいた報酬調整による，利他的な行動の自発的な学習を目指している．

3. タスク割当問題

3.1 タスク割当問題の定義

複数のメンバからなるコミュニティに対して複数のタスクが与えられており，各メンバに対する特性（得意なタスクや不得意なタスクなど）が定義されている環境を考える．リーダが存在するコミュニティでは，一般的にはリーダによるタスクの割当が効率的な方法であると思われる．しかし災害時における初期のボランティア活動のように，リーダが存在しない，あるいはリーダがメンバの特性を把握していないような環境では，リーダによる適切なタスクの割当は困難であり，コミュニティの各メンバが自身の処理するタスクを自発的に選択，担当する方が望ましい場合も考えられる．本論文では，このような状況のモデルとして下記に示すタスク割当問題を定義する．

Input: コミュニティに属するメンバの集合と各メンバの特性，複数種類のタスクからなるタスク集合，及びタスクを提示する順番．

Output: すべてのタスクを公平かつ迅速に処理するための，各メンバの行動規則．

また，タスク割当問題の詳細として以下を仮定する．

- タスクは一つずつ提示される．
- タスクを担当するメンバが決定するまでは，新しいタスクは提示されない．
- 各メンバは，現在提示されているタスクの種類，及び他のメンバの状態を知覚できる．
- 処理中のタスクをもたないメンバのみが，現在提示されているタスクを担当するか否かを選択できる．
- あるタスクの担当者は，そのタスクを希望するメンバの中から選定される．
- 各メンバには，得意なタスク及び不得意なタスクがある．
- 各メンバは，自分が得意とするタスク及び不得意なタスクを知らない．
- 不得意とするタスクを担当したメンバは，そのタスクを完了するために多くの処理時間を必要とする．一方得意とするタスクを担当した場合には，少ない処理時間でタスクを完了することができる．
- タスクを完了したメンバには正の報酬 R^w が与えられる．R^w はタスクの種類，及びタスクの完了に要した処理時間にかかわらず一定値とする．
- 各メンバの能力は変化しない．

本論文では，結果の解析を容易にするために上記のタスク割当問題を対象とする．しかし，複数のタスクを同時に提示する，タスクの完了に伴ってメンバの能力が向上するなどの拡張は容易である．

3.2 タスク割当問題の特徴

各メンバが，自分自身の受け取る報酬の最大化を目的として行動規則を学習する場合を考える．この場合，各メンバは学習を通じて自身の得意とするタスク及び不得意とするタスクを学習し，自分の得意とするタスクのみを処理するような行動規則を獲得する．しかし，すべてのメンバが不得意とするタスク（以下，ハードタスク）が存在する場合は，ハードタスクの処理を希望するメンバが存在しなくなり，タスクの割当が停滞することがある．この問題を解消するため，本研究では各メンバに対して勤勉性 IN_i を定義し，勤勉でないメンバに対して負の報酬 R^p をペナルティとして与えることにより，タスク割当の停滞を抑制する．なお，勤勉性に関する詳細については次章で述べる．

ペナルティの導入により，一部のメンバはハードタスクの処理を希望するような行動規則を獲得する．ハードタスクの処理を希望することは他のメンバの得る報酬の増加につながるため，利他的な行動と考えることもできる．しかしこの行動は，自分自身が受けるペナルティを最小化するための行動，すなわち利己的な行動ととらえた方が妥当であると思われる．また，一部のメンバのみがハードタスクを処理するという状況が発生しやすく，タスクを公平に処理するための行動規則の獲得は困難となる．

本論文では，ハードタスクを処理しない方が得策である場合でも，他のメンバあるいはコミュニティ全体の利益をもたらすためにハードタスクを希望する行動が利他的な行動であると考える．すなわち，すべてのメンバがハードタスクを希望する行動規則を獲得し，各メンバが受けるペナルティの回数が均等になるような行動規則が獲得されたとき，利他的な行動規則が獲得されたと考える．したがって本論文では，各メンバが処理するハードタスク数及び各メンバが受けるペナルティ回数が均等になるような行動規則の獲得をタスク割当問題の最も重要な目的とする．また利他的な行動規則を獲得する上での動機付けとして，他のメンバを思いやるなどの感情に着目し，感情が利他的な行動規則の獲得に及ぼす影響について検討する．

4. 強化学習による利他的な行動規則の獲得

本章では，タスク割当問題のマルチエージェントモデルを提示し，エージェントの行動規則の獲得手法について述べる．

4.1 タスク割当問題のマルチエージェントモデル

コミュニティに N 名のメンバ $A_i(i = 0, 1, ..., N-1)$ が存在すると仮定し，各メンバをそれぞれエージェントとして扱う．タスクの総数は M とし，提示される順番に $T_m(m = 0, 1, ..., M-1)$ と表記する．また，タスクは K 種類存在すると仮定し，タスクの種類を $T^k(k = 0, 1, ..., K-1)$ と表記する．以下に，本論文で仮定するタスク割当問題のシミュレーションモデルを示す．

1. 各エージェントの IN_i 及び行動規則を初期化する．また，m と t を 0 に初期化する．
2. T_{M-1} を処理するエージェントが決定するまで，下記を繰り返す．
   - 2.1. T_m を提示する．
   - 2.2. 各エージェントが知覚情報を受理する．
   - 2.3. 各エージェントが行動を決定する．
   - 2.4. T_m を希望するエージェントが存在する場合は，これを処理するエージェントを決定し，m を 1 増加する．
   - 2.5. IN_i を更新する．
   - 2.6. 各エージェントに報酬を与える．
   - 2.7. t を 1 増加する．

以降では，上記モデルにおける 2.1 から 2.7 までの処理を 1 ステップとして扱う．また，T_{M-1} を処理するエージェントが決定したときにエピソードが終了するものとし，そのときのステップ数（t）をエピソード長とする．なお，本モデルにおけるゴールは，下記の三つの値を最小化するようなエージェントの行動規則の獲得とする．

- エージェントがペナルティを受けた回数の総和．
- エージェントがペナルティを受けた回数の分散．
- エピソード長．

以降では，エージェントモデルの詳細について述べる．

4.2 エージェントの仕様

エージェントは現在提示されているタスク T_m の処理を希望するか否かの 2 種類の行動を選択できる．ただし，エージェントが一度に処理できるタスクは一つと制限されているため，現在遂行中のタスクをもたないエージェントのみが T_m の処理を希望できる．

各エージェントには，得意とする種類のタスク，不得意とするタスクが定義されている．得意とするタスクを処理するエージェントは少ないステップ数でこれを完了することができるが，不得意とするタスクを処理するエージェントは，その完了に多くのステップ数を必要とする．また，各エージェントは自分自身の特性（得意，不得意とするタスク）を知らないが，自分自身の特性は学習を通じて獲得するものと仮定する．なお，各エージェントがタスクを完了するために必要とするステップ数の詳細については **5.** で述べる．

各エージェント A_i に対して勤勉性 IN_i を定義する．IN_i は A_i が任意のタスクを処理するエージェントとして決定したときに増加し，かつシミュレーションステップ t の増加に伴って減少する．すなわち，IN_i が大きいエージェントは勤勉なエージェントとみなされ，IN_i が小さい場合は怠惰なエージェントとみなされる．

各エージェントは 2 種類の情報を知覚できる．一方は現在提示されているタスクの種類であり，他方は他エージェントの状態である．ここでエージェント A_i が知覚できる他エージェント A_j の状態は，以下に示す 3 状態のいずれかである．

- A_j がタスクを処理中である．
- A_j は処理中のタスクをもっておらず，A_j の勤勉性 IN_j が IN_i よりも小さい．
- A_j は処理中のタスクをもっておらず，IN_j が IN_i 以上である．

タスクの種類及びエージェント数はそれぞれ K，N と仮定しているため，各エージェントが知覚できる状態は $3^{(N-1)}K$ 通り存在する．

各エージェントは，Q 学習によって行動規則を学習する．Q 値の更新式を式 (1) に示す．ここで $Q_i(s_i^t, a_i^t)$ は A_i が参照する Q テーブル，s_i^t は t シミュレーションステップ時に A_i が受け取る知覚情報である．a_i^t 及び R_i^t は，t において A_i が選択した行動，及び A_i が受け取る報酬を表す．R_i^t の詳細については，**4.4** で述べる．なお，α 及び γ はそれぞれ学習率，割引率である．また各エージェントは，ε-greedy 戦略に基づいて行動を選択するものとする．

$$Q_i(s_i^t, a_i^t) = (1.0 - \alpha)Q_i(s_i^t, a_i^t) + \alpha(R_i^t + \gamma \max_{a_i^{t+1}}(Q_i(s_i^{t+1}, a_i^{t+1}))) \quad (1)$$

4.3 タスクを処理するエージェントの選定

タスク T_m を希望するエージェントが唯一であるときは，そのエージェントが T_m を処理するエージェントとして決定される．複数のエージェントが T_m を希望する場合には，これらの中で IN_i が最も小さいエージェントが選定される．なお，IN_i が最小であるエージェントが複数存在する場合は，これらの中で，最後にタスクを割り当てられてから現在までのステップ数が最も大きいエージェントがランダムに選択される．T_m を希望するエージェントが存在しない場合は，システムはシミュレーションステップ t を 1 増加し，T_m を希望するエージェントが現れるまで T_m の提示を継続する．

4.4 報酬と感情

A_i が受け取る報酬 R_i は正の報酬 R_i^+ と負の報酬 R_i^- の総和として計算される．R_i^+ がタスクを完了することによって与えられる報酬 R^w と等しく，かつ R_i^- が A_i の勤勉性が低いことによって与えられるペナルティ R^p と等しい場合は，**3.** で述べたように利他的な行動規則の獲得が困難となる．したがって，本論文では 2 種類の感情を導入し，これを報酬に反映させることによって利他的な行動規則の獲得を試みる．

1 番目の感情は，自分以外のエージェントがペナルティを受けたときに，そのエージェントに対して同情する感情（以下，憐憫悲）である．ある時点においてペナルティを受けるエージェント数が S である場合を仮定すると，ペナルティを受けるエージェントに対する負の報酬は $R_i^- = R^p + (S-1)R^s$ として，ペナルティを受けていないエージェントに対する負の報酬は $R_i^- = SR^s$ として定義する．ここで $R^s \leq 0$ は憐憫悲の大きさを表すパラメータである．

2 番目の感情は，自分以外のエージェントがタスクを完了し正の報酬を受け取ることを喜ぶ感情（以下，憐憫喜）である．憐憫喜を導入する場合の正の報酬は，憐憫悲と同様に計算される．すなわち，タスクを完了したエージェント数を G，憐憫喜の大きさを表すパラメータを $R^g \geq 0$ と表記すると，タスクを完了したエージェントの受け取る正の報酬は $R_i^+ = R^w + (G-1)R^g$ として，タスクの完了による報酬を受け取らないエージェントの正の報酬は $R_i^+ = G\,R^g$ として計算される．

憐憫悲，憐憫喜以外の感情を報酬に反映することは可能である．例えば，自分以外のエージェントがペナルティを受けることを喜ぶ感情（嘲笑）は $R^s > 0$ を仮定することにより，自分以外のエージェントが正の報酬を受け取ることを悲しむ感情（嫉妬）は $R^g < 0$ を仮定することにより，報酬に対する影響を表現することができる．しかしこれらの感情の導入は，エージェントの受けるペナルティ回数の均等化，並びにハードタスクの公平な分担の妨げになることが予想される．このため本論文では，憐憫悲及び憐憫喜による影響のみについて検討する．

5. 実験及び検討

5.1 実験パラメータ

4. で述べたマルチエージェントモデルを Sun Ultra10(333 MHz) 上に C 言語により実装し，シミュレーション実験を行った．実験では，エージェント数 N を 3，タスクの種類 K を 4，タスクの総数 M を 1000 とした．各エージェント及びタスクの特徴を表 1 にまとめる．この表は，エージェントがタスクの処理に要するステップ数を表している．T^3 はすべてのエージェントが不得意とするタスク，すなわちハードタスクであり，T^3 の処理を希望することが利他的な行動となる．また，解析を容易にするためにタスクの発生順序を固定し，$T_{4n} = T^0, T_{4n+1} = T^1, T_{4n+2} = T^2$，$T_{4n+3} = T^3 (n = 0, 1, ..., 249)$ とした．

タスクを完了したことによる正の報酬 R^w は 0.15，怠惰なエージェントに対して与えられる負の報酬 R^p は -0.03 とした．また，エージェントの勤勉性 IN_i の初期値は 6 とした．IN_i はステップごとに更新され，現在提示されているタスクの担当者となったエージェントに対しては上限値を 9 として IN_i を 3 増加させ，その他のエージェントについては下限値を 0 として IN_i を 1 減少させた．なお，各ステップにおいてペナルティを受けるエージェントは，その時点において IN_i が 0 であるすべてのエージェントとした．IN_i の更新方法，及び表 1 に示す値は，得意とするタスク（処理に要するステップ数が 1 であるタスク）を処理するエージェントの勤勉性は増加し，不得意とするタスク（処理に要するステップ数が 5 であるタスク）を処理するエージェントの勤勉性は減少するように設定されている．

表 1 タスクの完了に要するステップ数
Table 1 The number of steps to complete a task.

	T^0	T^1	T^2	T^3
A_0	1	3	5	5
A_1	5	1	3	5
A_2	3	5	1	5

学習率 α，割引率 γ はそれぞれ 0.9，0.8 とし，行動の選択には ε-greedy 方策を使用した．一様分布を用いて行動を選択する確率 $\varepsilon(\%)$ は式 (2) によって更新するものとし，エピソード数の増加に伴って単調減少するように設定した．なお max_e 及び e は，それぞれエピソード数の最大値，現在のエピソード数を表している．また，$\lceil x \rceil$ は x を超えない最大の整数である．

$$\varepsilon = \left\lceil \frac{max\_e - e}{3} \right\rceil \tag{2}$$

なお本論文では，エピソード数を 100 としたときの学習を 1 試行とし，これを 1000 試行実施したときの平均によって結果を評価する．

5.2 実験結果

はじめに，感情による影響を考えない場合の結果を提示する．得られた行動規則の約 95%は表 2 に示す 9 パターンのいずれかであった．表中の LE は各パターンにおけるエピソード長，NP は各エージェントがペナルティを受け取った回数，$\#T^3$ は各エージェントがハードタスク T^3 を処理した回数を表している．これらの値は，学習終了時における行動規則を使用したときの結果である．また，各パターンの出現頻度を図 1 に示す．感情による影響を考えない場合では，エピソード長が長く，かつ NP の偏りが大きいパターンである α 及び β の出現頻度が支配的であった．ペナルティを受けた回数の平均は，A_0 が 139 回，A_1 が 324 回，A_2 が 310 回であり，A_0 がペナルティを受ける回数が他のエージェントに比べて著しく少ない結果となった．また A_0 が T^3 を処理した回数も，パターン η を除いて他のエージェントよりも少ない結果が得られた．タスクを提示する順番が A_0 にとって有利に働いたことが，このような結果が得られた原因と考えられる．

例えば $t = c$ において A_0 が T^3 を処理するエージェントとして選定された場合を考える（図 2）．図中の長方形は各エージェントがタスクを処理している期間を表している．この場合，A_0 が T^3 の処理を終了したとき（$t = c+5$）に提示されているタスクは，T^0 あるいは T^3 である可能性が高い．A_0 が得意としている T^0 が提示されている場合は，A_0 がこれを処理することにより，勤勉性 IN_0 を回復（増加）することが可能である．更に，T^0 の処理が終了した後に A_0 が T^1 を処理するエージェントとして選定されると，その後に提示される T^2 及び T^3 の処理を回避できる確率が高くなる．T^3 が提示されている場合でも，T^0 が提示される（他のエージェントが T^3 の担当エージェントとして決定する）まで待機することにより，IN_0 を増加させることが可能である．一方 A_1 が T^3 を処理する場合（図 3）では，A_1 が T^3 の処理を完了した後に提示されているタスクは，A_1 が不得意としている T^0 あるいは T^3 である可能性が高い．このため A_1 は，T^3 の完了後に IN_1 を回復させることが困難となり，その後高い確率でペナルティを受けることになる．A_2 についても同様に，T^3 の完了後に IN_2 を

表 2　獲得された行動規則の特徴
Table 2　Characteristics of acquired rules.

Pattern	LE	NP			$\#T^3$		
		A_0	A_1	A_2	A_0	A_1	A_2
α	1250	120	120	490	0	125	125
β	1373	240	620	240	60	125	65
γ	1126	0	120	240	40	85	125
δ	1102	50	50	200	50	100	100
ε	1207	80	330	200	40	85	125
ζ	1200	150	300	150	50	100	100
η	1125	0	0	360	125	0	125
θ	1250	120	490	120	0	125	125
ι	1332	80	580	330	0	85	165

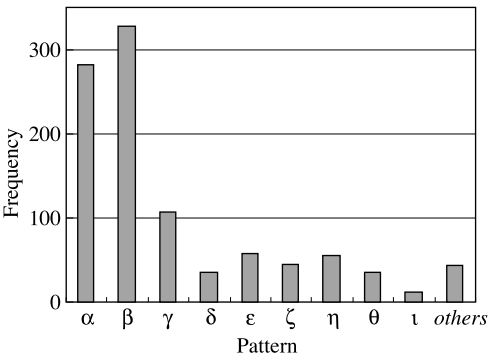

図 1　各パターンの度数分布
Fig. 1　Frequency of each pattern.

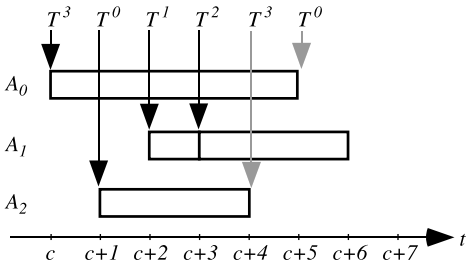

図 2　タスクの割当（A_0 が T^3 を処理する場合）
Fig. 2　An example of task allocation when T^3 is assigned to A_0.

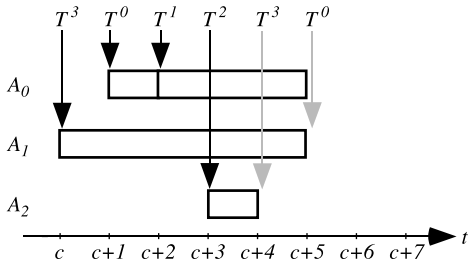

図 3　タスクの割当（A_1 が T^3 を処理する場合）
Fig. 3　An example of task allocation when T^3 is assigned to A_1.

回復させることが A_0 に比べて困難である．このように感情による影響を考えない場合では，各エージェントは自分の受け取る報酬の最大化のみを目的とするために利他的な行動規則の獲得が困難となり，エージェントが受けるペナルティの回数に偏りが生じる傾向が見られる．

次に，憐憫悲を報酬に反映したときの結果を提示する．図 4 及び図 5 は，他エージェントがペナルティを受け取ったときに発生する負の報酬 R^s を 0 から -0.03 まで変化させたときの結果である．図 4 は平均エピソード長（LE），図 5 は各エージェントがペナルティを受けた平均回数（NP）を表している．R^s の減少（憐憫悲の報酬に対する影響の増加）に伴い，LE 及び NP は減少している．また，R^s の減少に伴って，A_1 が受けるペナルティの回数が著しく減少していることから，各エージェントが受けたペナルティの分散も減少傾向にある．図 6 に，得られた行動規則パターンの分布を示す．R^s の減少に伴い，A_2 が多くのペナルティを受けるパターンである α，及び A_1 が多くのペナルティを受けるパターンである β が減少している．すなわち，憐憫悲の導入により，各エージェントが受けるペナルティの分散が減少する傾向が見られた．また，A_0 が多くの T^3 を処理するパターンである η が著しく増加していることから，憐憫悲の導入によって T^3 が公平に分担される傾向も見られた．これらのことより，憐憫悲の導入は利他的な行動規則の獲得に寄与していると考えられる．

次に，過度の憐憫悲を導入したときの結果を提示する．図 7 は，R^s を -0.03 に固定し，ペナルティによる負の報酬 R^p を -0.03 から 0 へ変化させたときの結果を表している．憐憫悲による影響 R^s がペナルティ R^p よりも大きい場合は，R^p の増加（R^p の報酬に対する影響の減少）に伴い，エージェントが受けるペナルティの総数が単調増加している．この結果から，憐憫悲を適切な大きさで報酬に反映することは利他的な行動規則の獲得に貢献するが，過度の憐憫悲の導入は利他的な行動規則の学習においては障害となることが示された．

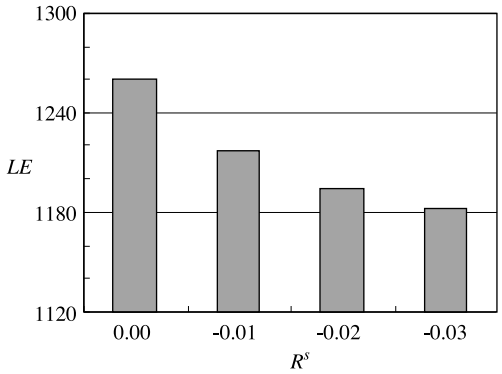

図 4　憐憫悲（$R^p \leq R^s \leq 0$）の導入による LE の変化
Fig. 4　The length of an episode when $R^p \leq R^s \leq 0$.

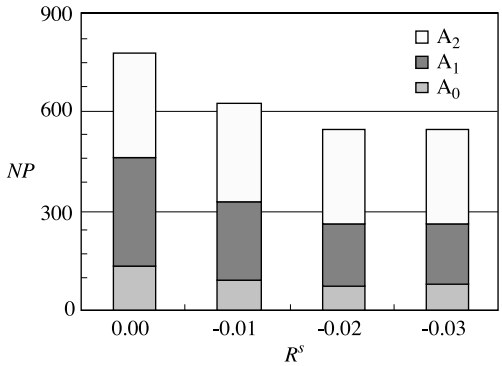

図 5　憐憫悲（$R^p \leq R^s \leq 0$）の導入による NP の変化
Fig. 5　The number of times that agents receive penalty when $R^p \leq R^s \leq 0$.

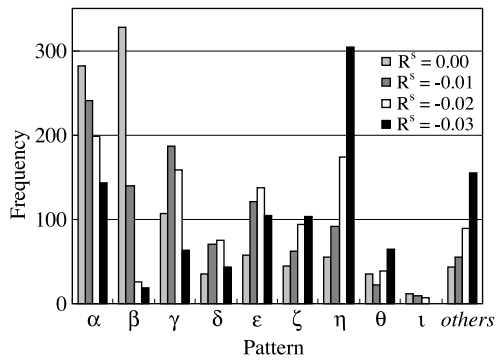

図 6　各パターンの分布の変化
Fig. 6　Changes of frequency of each pattern.

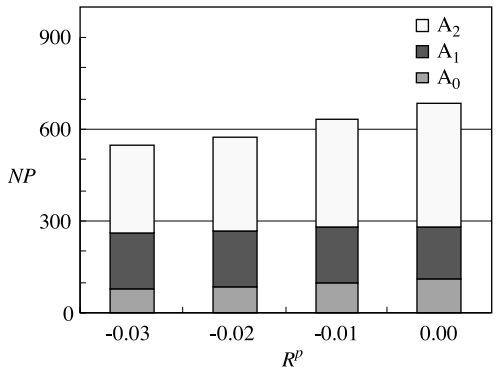

図 7　憐憫悲（$R^s \leq R^p \leq 0$）の導入による NP の変化
Fig. 7　The number of times that agents receive penalty when $R^s \leq R^p \leq 0$.

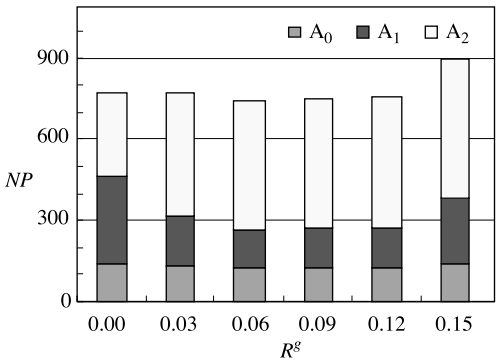

図 8　憐憫喜（$0 \leq R^g \leq R^w$）の導入による NP の変化
Fig. 8　The number of times that agents receive penalty when $0 \leq R^g \leq R^w$.

最後に憐憫喜を報酬に反映したときの結果を提示する．図 8 は，任意のエージェントがタスクを達成したときに他のエージェントに対して発生する正の報酬 R^g を 0 から 0.15 まで変化させたときの結果を表している．R^p 及び R^s はそれぞれ -0.03，0 に固定している．$R^g < R^w$ である場合は，R^g の増加に伴ってエージェントがペナルティを受けた回数（NP）の総数がわずかに減少しているが，NP の分散は大きくなる傾向にあった．また $R^g = R^w = 0.15$ である場合は，憐憫喜を導入しない場合に比べてペナルティの総数，及びその分散が増加傾向にある．これらのことより，憐憫喜の導入は，利他的な行動規則の学習に寄与しないことが示された．

5.3 検　　討

タスク割当問題に対する実験結果から，憐憫悲を適切な大きさで報酬に反映することは利他的な行動の動機付けとなることが示された．しかし本論文で提示した結果は，ハードタスクの処理に要するステップ数，タスクを完了したことによる正の報酬 R^w，怠惰なエージェントに対して与えられる負の報酬 R^p などのパラメータに強く依存している．このため，憐憫悲を報酬に反映する際の適切な値は，パラメータ値の変更に伴って変動することが予想される．憐憫悲を報酬に反映する際の最適値をパラメータ値から推定するためには，より詳細な解析が必要であると考えている．

本論文で提示した感情を導入した強化学習法の，タスク割当問題以外の問題に対する適用も検討すべき課題である．追跡問題 [13] のような，その他の競合状況に分類される問題や，互恵的競合状況に分類される他の問題に対して提案手法を適用し，感情が利他的な行動の選択に及ぼす影響について検討すべきであろう．なお，互恵的競合状況に分類される問題としては，狭路問題 [12] を拡張した問題について現在検討中である．狭路問題は，2 台の車（エージェント）が同時にすれ違うことのできない狭路の両端に対峙している状況を想定している．この問題を複数台の車が対峙する問題，すなわち交通事故や道路工事などで実施される片側交互通行のような問題へと拡張し，事故現場（工事現場）を通過することへの正の報酬，相手を待つことへの負の報酬に加え，自分が前に進めないこと，及び前進したことによる報酬を導入することにより，互恵的競合状況に分類される問題のモデル化を試みている．

最後に，サッカーゲームや会社組織などの実社会では，集団を統括するリーダなどが存在し，報酬やタスクを適切に配分することにより集団全体での利益の向上がなされている．しかしこのような環境においても，社会の構成員には感情があり，これが集団全体の利益に影響を及ぼしていると考えられる．集団を統括するリーダなどが存在するような環境，及び報酬の配分機構 [11], [14] が実装された環境における，感情の影響について検討することも今後の課題である．

6. む　す　び

本論文では，人間社会の簡単なモデルとしてタスク割当問題を定義し，そのマルチエージェントモデルを提示した．また，利他的な行動規則を獲得するための方策として 2 種類の感情を導入した強化学習手法を提示した．計算機実験による結果からは，適度な大きさの憐憫悲の導入は利他的な行動規則の獲得に寄与することが示されたが，憐憫喜及び過度の憐憫悲の導入

は利他的な行動規則の獲得には寄与しないことが示された．

本論文では，解析を容易にするために単純な問題を対象とした．また，感情の利他的な行動規則の獲得への影響について論じるため，検討した感情も憐憫悲及び憐憫喜の 2 種類であった．しかし，現実の人間社会は非常に複雑であり，嫉妬や嘲笑などの様々な感情が，社会の構成員の意思決定に影響を及ぼしている．より複雑な社会モデルを構築し，様々な感情が意思決定に及ぼす影響を検討することが今後の課題である．

謝辞 貴重なコメントを頂いた担当編集委員並びに査読委員の方々に感謝致します．本研究の一部は，広島市立大学特定研究費の助成による．

文　献

[1] http://www.robocup.org
[2] H. Kawamura, A. Ohuchi, and K. Kurumatani, “Development of X-economy system for simulation of multi-agent economy,” in Agent-Based Approaches in Economic and Social Complex Systems, ed. A. Namatame, T. Terano, and K. Kurumatani, pp.188–197, IOS Press, 2002.
[3] 谷本　潤，藤井晴行，“マルチエージェントシミュレーションによる談合モデル，” 情処学研報，2003-ICS-131-12, pp.63–68, Jan. 2003.
[4] S.J. Russel and P. Norvig, Artificial Intelligence: A Modern Approach, Prentice-Hall International, 1995.
[5] B. Mesot, E. Sanchez, C-A. Peña, and A. Perez-Uribe, “SOS++: Finding smart behaviors using learning and evolution,” Proc. Artificial Life, VIII, pp.264–273, Dec. 2002.
[6] 鈴木麗璽，有田隆也，“進化と学習の相互作用—繰り返し囚人のジレンマゲームにおける Baldwin 効果，” 人工知能誌，vol.15, no.3, pp.495–502, 2000.
[7] H. Katagiri, K. Hirakawa, and J. Hu, “Genetic network programming—Application to intelligent agents,” Proc. IEEE International Conference on System, Man and Cybernetics, pp.3829–3834, 2000.
[8] 平澤宏太郎，大久保雅文，片桐広伸，胡　敬炉，村田純一，“蟻の行動進化における Genetic Network Programming と Genetic Programming の性能比較，” 電学論（C），vol.121-C, no.6, pp.1001–1009, June 2001.
[9] H. Ueda, N. Iwane, K. Takahashi, and T. Miyahara, “Acquisition of a state transition graph using genetic network programming techniques,” Proc. TENCON2003, pp.163–167, Oct. 2003.
[10] H. Ueda, T. Tanizawa, K. Takahashi, and T. Miyahara, “Acquisition of reciprocal altruism in a multi-agent system,” Proc. TENCON2004, pp.334–337, Nov. 2004.
[11] 西岡靖之，藤田敏之，“利得制御による自律分散型エージェントの協調関係の実現，” 人工知能誌，vol.15, no.6, pp.1043–1051, 2000.
[12] 森山甲一，沼尾正行，“環境状況に応じて自己の報酬を操作する学習エージェントの構築，” 人工知能誌，vol.17, no.6, pp.676–683, 2002.
[13] 渡邉亮介，丸山文宏，永田裕一，東条　敏，“互恵性原理に基づくマルチエージェントの強化学習法，” 第 16 回人工知能学会全国大会論文集，2D3-01, 2002.
[14] 保知良暢，新谷虎松，伊藤孝行，大囿忠親，“ 外部評価機構を導入したマルチエージェント強化学習における過去の事象に基づく報酬配分，” 信学論（D-I），vol.J87-D-I, no.12, pp.1119–1127, Dec. 2004.

（平成 16 年 11 月 4 日受付，17 年 2 月 23 日再受付）

上田　祐彰　（正員）

平 2 広島大・総合科学・総合科学卒．平 4 同大大学院工学研究科博士課程前期了．平 7 阪大大学院工学研究科博士後期課程単位取得退学．現在，広島市立大学情報科学部助手．博士（工学）．遺伝的アルゴリズム，機械学習等に関心をもつ．情報処理学会，人工知能学会，IEEE 各会員．

谷澤　俊彰

平 16 広島市大・情報科学・知能情報システム工卒．現在，京大大学院情報学研究科博士前期課程在学中．利他・共生システム，マルチエージェントシステム等に関心をもつ．

高橋　健一　（正員）

昭 52 名工大・工・情報卒．昭 54 同大大学院修士課程了．同年同大工学部情報工学科助手．昭 62 同大電気情報工学科講師．平元同大助教授．平 6 広島市立大学情報科学部教授．工博．機械学習，画像処理に関する研究に従事．情報処理学会，人工知能学会，IEEE 等各会員．

宮原　哲浩　（正員）

昭 59 九大・理・数学卒．昭 61 同大大学院・総合理工学・情報システム学修士課程了．昭 63 同博士課程中退．昭 63〜平 8 九大・理学部及び教養部に勤務．平 8 広島市立大学情報科学部助教授．博士（理学）．主に機械学習の研究に従事．情報処理学会，人工知能学会各会員．