計測自動制御学会論文集
Vol.39, No.5, 1/9（2003）

強化学習による個性・社会性の発現・分化モデル †

柴田克成*・上田雅英**・伊藤宏司***

Emergence and Differentiation Model of Individuality and Sociality by Reinforcement Learning

Katsunari Shibata*, Masahide Ueda** and Koji Ito***

In this paper, a concept of individuality and sociality is introduced as a method to avoid conflicts of individual interests in multi-agent systems. It is considered that each agent has its individuality when the conflicts are resolved by making its own mapping from the sensory input to the action output. On the other hand, each agent has sociality when the conflicts are avoided by some common input-output mapping, which is commonly called rules. A conflict avoidance task in which passengers are getting on and off a train are taken as an example, and the emergence processes of both behavioral characters are explained. Furthermore, it is shown that the differentiation of the agent into one of them is adaptively realized by reinforcement learning based on local rewards according to the asymmetry of environment, number of agents, identification of the other agents, or physical ability of agents.

Key Words: Individuality, Sociality, Conflicts, Multi-Agent System, Reinforcement Learning

1. はじめに

1.1 利害の衝突回避のための個性と社会性

人間社会において，個々の人々は様々な個性を持っている反面，社会のルールに従う社会性という側面も持ち合わせている．この一見相反する個性と社会性はどうして存在するのか，またどのように形成されるのであろうか？

たとえば，集団の中での意志決定を考えてみる．もし，全員が自己の利益を追求し，意志を曲げなければ，その調整のために時間を費やし，結局全員が不利益を被ることになる．このような時に，自分の意志を曲げない “わがまま” という個性を持った人と “優柔不断” という個性を持った人がいれば，” 優柔不断” な人が譲歩することで，意志決定はよりスムーズに進むと考えられる．このように，集団での利害の衝突回避というものが，個性の発現の一つの要因になっていると考えることができる．

一方，たとえば，交差点での直進車と反対車線の右折車は，直進車が優先するという交通ルールによって，事故を避け，スムーズな交通の流れを実現している．このことから，ルールに従う社会性というものの発現も，利害の衝突回避が一つの要因であると考えることができる．つまり，個性と社会性は，一見相反する特性のように見えるが，どちらも利害の衝突回避を発現要因の一つと考えることができる．

利害の衝突回避は，マルチエージェント系で効率的な動作を望む場合，必要不可欠なものである[1)]．相手が存在しているにもかかわらず，相手が存在しないときの最適な戦略をとっていると，デッドロックに陥ったりして，結果として両者共に利益が減少するという状況になってしまう．分散人工知能の分野では，1 対 1 の競合問題を扱ったゲーム理論の研究がなされている．ここでの分類[1)]に従うと，本研究で扱う問題は，エージェント間の通信を考えておらず，さらに，一方の利得が他方の損失を意味するわけではないため，非協力非ゼロサムゲームに分類される．したがって，1 対 1 の場合を考えると，それぞれのエージェントごとに利得行列を考えることができる．しかし，1 回の行動決定だけで報酬が決まるわけではなく，その後の行動も学習によって変化するので，利得行列自体が時間とともに変化することになる．また，多対多の場合も考えていかなければならない．

1.2 個性・社会性に関する従来研究

従来，個性というものは，先天的に備わっているものとする見方が強く，有名なハト・タカゲームでは，ハト派とタカ派という個性を予め与えた上で，その存在比率に関して論

† 第 11 回自律分散システムシンポジウムにて一部発表 (1999.1)
* 大分大学 工学部 電気電子工学科
** (株) 本田技術研究所栃木研究所
*** 東京工業大学大学院総合理工学研究科知能システム科学専攻
* Dept. of Electrical & Electronic Engineering, Faculty of Engineering, Oita Univ.
** Honda R&D Co., Ltd. Tochigi R&D Center
*** Dept. of Comp. Intel. & Sys. Sci., Tokyo Inst. of Tech.
（Received November 19, 2001）
（Revised August 5, 2002）

じている[2]．また，マルチエージェントシステムや群ロボットの研究では，個性を付与したロボットの混合比率に対するシステム全体としての能力を調べたり[3]，「利他的・利己的」などの予め用意された複数の個性の中から，学習を通して一つの個性を獲得させ，それが集団に与える影響を論じている研究もある[4) 5)]．また，社会学の分野では，ジンメルは，個人が複数の社会に同時に属することによって特定の社会の拘束からはずれ，個性が生ずるとしている[6]．つまり，個人と社会の関係の中から個性について論じているが，なぜ社会が形成されるかといった点も含めて，個性が生ずるまでの過程について，個体のアルゴリズムが明らかになっているとは言えない．また，生態学の分野では，社会の中で順位が確立されると，争いに消費されるエネルギーは集団の総合的機能のために用いられると論じているものもある[7]．この考え方は，本研究の個性の発現に対する考え方に近いものである．

一方，社会性やルールに関しては，社会というものに元々備わっているものとする見方が強い．たとえば，ホッブスは，社会的ジレンマを解くためにはそれを罰する権力が必要と考え，国家の必要性を主張している[8]．社会的制裁機構の存在により，人々は規則に従う行動をとるようになるが，規則は権力，つまり，上から与えられるものとして考えられている．しかし，社会は元々，個体の集合であり，ルールを社会が元来もっているとするのには無理があると筆者らは考える．また，組織の中で，個体と知識共有の場である組織場とのインタラクションを通して学習を行う組織学習という考え方もある[9]が，やはり，組織場というものの存在が前提となっている．このような中で，太田らは，群ロボットの経路選択問題においてに，学習によって環境に応じた経路選択ルールが発現することを示した[10]．ただし，ここでは，個々のロボットが学習するのではなく，群としての戦略を逐次変化させていく形であり，また，学習も，強化学習的ではあるが，一般的な遅延報酬の形態にはなっていない．

1.3 本研究の目的

本研究では，社会や組織というものは，個にとって利益をもたらすものであり，個が自己の利益を追求する学習を行った結果生じるものと考える．また，個の個性というものは，その行動を外部から観測し，他の個と比較することによって判断されるものであると考える．したがって，社会も個性も予め存在するものではなく，個が単に自己の利益の追求のために行動を学習することによって，多数の個が集まった集団，つまりマルチエージェントシステムとして見たときに，個性や社会性として外部から観察できると考える．これらの考え方に基づいて，個性・社会性の新たな発現モデルを提唱することを本研究の一つの目的とする．そして，ゆくゆくは人間や生物の社会における個性・社会性形成のモデルとして発展し，社会の解析に役立つことも期待している．

また，本研究は，人間の社会行動を含む生物の高次機能の解明とそれをロボットに実装するという大きな目標に向けた研究の一環でもある．筆者らは，強化学習は単に行動だけでなく，認識，注意[11]，記憶，制御，会話[12]などにも有効であり，これらを強化学習に基づいて調和的に学習させることが人間や生物の知能の源であり，主に知識を付与される形で知能化されている現在の知能ロボットと生物の大きな溝を埋めるキーとなると考えている[13) 14) 15)]．そのため，強化学習が，個性・社会性といった社会の中の特性を発現させる能力を持っていることを示すことを目的とするとともに，人間や生物が強化学習を行っている可能性を示す例としてもとらえる．

以上より，本研究では，獲得された個性・社会性およびその分化に対する評価は，いかに人間の挙動に近いか，いかに人間が理解できるかというところがよりどころとなる．そして，本論文では，問題自体，および人間がどう行動するかが一般的に理解しやすい問題として電車の乗降問題を例題として取り上げた．また，多くの従来研究のように，予めエージェントに何らかの特性を与えて，それを学習させる場合と異なり，単に行動を学習させるため，定量的な評価は難しいが，設計者の意図が入りにくく，学習の自由度も大きい．そして，学習によって得られた結果に対してのみ，外部観測者として筆者らが解釈を加えることによって評価を行っていく．また，本研究では，特に新しい手法を提案しているわけではないが，個性・社会性発現のために特化した手法を提案するよりも，強化学習で行動を学習するだけで個性・社会性の発現と分化が起こることを示す方が，学習が容易であり，他の機能の学習との調和を考えても有効であると筆者らは考える．

本稿では，まず，個性と社会性を定義する．そして，われわれの社会を振り返り，環境の違いなどの個性と社会性が分化する要因を考える．つぎに，電車の乗り降り問題を例として，個々のエージェントが遅延報酬を前提とした強化学習を行い，その結果，個性・社会性を獲得できることを示す．さらに，例題の中に前述の分化要因を導入してそれを変化させることによって，同一の学習則に基づいて，適応的に個性と社会性に分化することを示し，その分化がわれわれの社会でのイメージと一致することを確認する．

2. 個性と社会性

2.1 定義

はじめに，ここで扱う個性と社会性に関して以下のように定義を行う．

- **個性 …** Fig.1 (a) のように，入力が同じでも，個体間に異なった行動を出力すること．すなわち，感覚入力－行動のマッピングが個体ごとに異なること．
- **社会性 …** Fig.1 (b) のように，感覚入力－行動のマッピングは個体間で同じであるが，感覚入力の違いによってのみエージェントの行動が異なること．

先に挙げた例を用いて説明すると，集団の意思決定を行う場合に，皆が「全体の意見が分かれてしまっている」という同じ入力を得ているにもかかわらず，「リーダーシップを

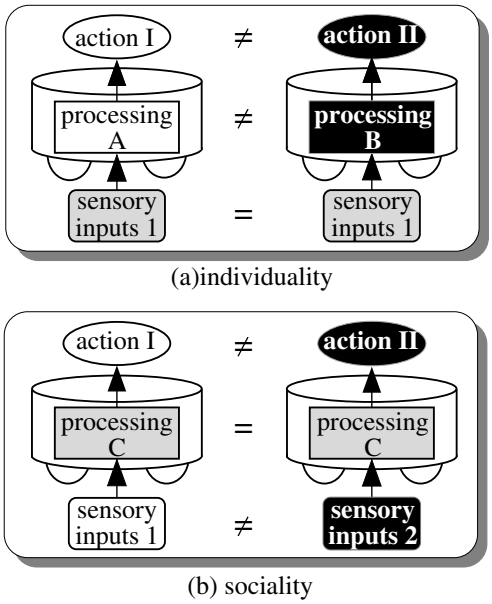


Fig. 1　Definition of individuality and sociality.

取る」者と「他者に同調する」者というように行動出力が異なっているのが個性である．一方，交差点で直進車と右折車が出会った場合に，どの車であろうと,「自車：直進，相手車：右折」の入力に対しては「進む」行動を出力し,「自車：右折，相手車：直進」の入力に対しては，相手車に進路を「譲る」行動を出力するというように，入力が同じ時には同じ行動を，入力が異なる時に異なる行動をとることが社会性である．これをマルチエージェントシステム全体から見たときには,「ルールが形成された」と呼ぶことができる．

社会性というと,「社会性を育てる」や「社会的規範」という時のルールに従うという意味と，ハチやアリなどの「社会性昆虫」という場合の集団での役割分担，さらに，social robot のように，インタラクションという意味あいで使われることもある．ここでは，前述の定義からわかるように，最初の，つまりルールに従うという意味で社会性という言葉を用い，集団での役割分担などは，逆に個性としてとらえる．ただし，一般に，社会性昆虫の集団中での役割は，遺伝的に決まっているが，本論文では，学習によって獲得されるものを扱う．

本論文では，個性と社会性というものが各エージェントに元来備わっているものとしていないため，外部の観測者がエージェントの行動を観察し，他のエージェントとの関係から，個性が発現したか社会性が発現したかを判断する．Fig.2 で格子状の簡単な環境での例を示す．2つのエージェントが向かい合って，お互いに進路を妨げているとする．そして，片方が譲歩したとする．ここで，両者の立場を入れ換えた時，(a) のように，同じエージェントが常に譲歩する場合，これを個性と判断する．一方，(b) のように場所を入れ替えた場合に，同じ場所のエージェントが同じ行動をすれば，これを社会性と判断する．この場合，学習後には確率的

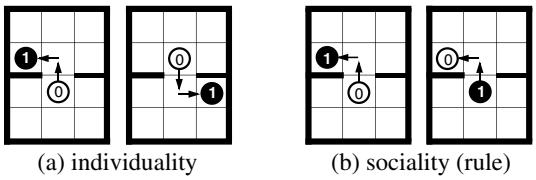


Fig. 2　Individuality and sociality in a simple example.

要素がなくなると考えれば，入れ替わることによって両エージェントの感覚入力に変化があったことになる．

2.2　3 つ以上の個体での定義

前述のように，1 対 1 の場合は，両者を入れ換えてその行動の差を観察することにより，個性・社会性の評価を行うことができる．個体数が 3 つ以上になった場合は，ある特定の個体との関係を，その立場を入れ替えることによって，1 対 1 の場合と同様に個性と社会性が定義できる．その上で，さらに他のすべての個体との関係から，個性による解決が多い場合は個性的な個体，社会性による解決が多く見られた場合は社会的な個体と定義する．したがって，個体数が増えるにしたがって，個性的または社会的であることを示す指標が連続値に近づいていく．

2.3　「利己的な個」の前提

経済学の観点でアダム・スミスは，個々が利己的であっても，市場は需要と供給のバランスで自動的に調節され，公共の福祉にかなういわゆる「見えざる手」による社会調和を主張している [16]．また，リチャード・ドーキンスは，われわれ生物の利他的行動は，実は遺伝子の利己的な性質に基づくとしている [17]．本研究でも，これらの考え方に近く，個体の行動原理に利他的なものは一切なく，すべて利己的なものであるとする．そして，あくまでも，個体が利己的に学習した結果，個性・社会性が発現し，この中には利他的と解釈できる行動が発現したりして，社会全体から見て調和するものとボトムアップ的に考える．

2.4　個性と社会性の分化の仮説とその要因

個性と社会性は，一見相反するものであるが，利害の衝突回避という共通の目的のために，状況に応じて適宜使い分けられている，つまり，学習によって分化するものと考える．

われわれ生物の社会を振り返ってみたときに，分化の要因としては以下のようなものが挙げられる．

● 環境の対称性

環境が完全に対称である場合には，相手のエージェントと立場を入れ換えても、個体が得る感覚入力に差異がない．したがって，利害の衝突回避を行うためには，個性によるしかない．非対称である場合には，立場を入れ換えると得られる感覚入力に差異があるので，その差異を利用して異なる行動を出力する社会性によって利害の衝突を回避することも可能である．ただし、環境の違いが学習結果にどう反映されるかはタスクや各個体の特性などによって変化するべきものであるため，定量化することは困難である．

● エージェント数と相手の特定

エージェント数が多くなれば，自分と同じ個性を持つエージェントが存在する可能性が多くなり，結果として利害が衝突する可能性が大きくなる．さらに，1 対 1 の単純な関係ではなくなる場合が多いため，個性を発揮して衝突を回避することは困難であると考えられる．したがって，エージェント数が多くなるほど社会性が発達すると考えられる．また，3 つ以上のエージェントが存在する場合，相手が特定できるかどうかが要因として考えられる．もし相手が特定できなければ，相手がどのような個性を持った個体であるかが判別つかないことになる．したがって，何らかのルールを形成する方が一般的に利害の衝突回避が容易になるため，社会性が発現しやすいと考えられる．一方，相手が特定できれば，その相手の個性によって自己の行動を変化させることが可能であり，個性による利害衝突回避が容易となり，個性が発現しやすいと考えられる．ただし，外見から相手が特定できなくても，相手の行動の履歴やコミュニケーションを通して相手の個性を判断することが可能である．しかし，本論文では，そこまでは考慮しない．

● 身体的能力差

身体的能力に大きな差があるエージェントが存在した場合，ひとくくりに同じルールでその行動を規定することは困難である．また，たとえそれができたとしても，身体能力の高い個体が存在すると，その個体にとっては，個性を発揮した方がより報酬を得やすくなる可能性が高く，社会性が発現しにくいと考えられる．身体的能力差は，それ自身が個性と考えることができるが，ここでは，定義に基づき，あくまでも個性は行動の差としてとらえる．また，身体的能力差が大きい場合は，同じ行動を取ること自体不可能な場合もある．このような場合は，先天的な個性と考えることもできるが，本論文中では取り扱わない．

後述のシミュレーションでは，提案するモデルによって，上記のような分化要因を変化させて，適応的に個性と社会性に分化するかどうかを確認し，強化学習に基づく個性・社会性の形成の有効性を検証する．

ここで，実世界における上記の分化要因について考えてみる．まず，環境が全く対称であるということはあり得ない．このことから考えると，社会性が有用であると考えられる．しかし，各個体の身体能力はかなりばらついており，身体能力が優位なものは，環境が非対称であっても，個性を発揮した方が得であり，これが個性を発現させる要因になっていると考えることができる．また，都会での電車の乗り降りのように，不特定多数での行動は，降車優先のルールがあることからわかるように，社会性が発現しやすいこともわかる．ただし，中にはルールを守らない者もいるし，それが必ずしも身体的能力差と言えない場合もある．これは，エージェントの内的な要因であると考えられ，遺伝によるもの，および，それまでの個体の経験による学習に基づくものがあると考えられる．しかし，このような問題は本論文では扱わない．

3. 利害の衝突の分類

マルチエージェントシステムにおいて，利害の衝突はさまざまな形で起こる．ここでは，利害の衝突を，その強さ，つまり，相手に譲歩することによってどれだけもらえる報酬が減少するかで分類する．従来の方法に基づき，エージェントの状態は，それ以後に得られる報酬の重みつき総和として次式のように評価される．

$$V_t = \sum_{i=1}^{\infty} \gamma^{i-1} r_{t+i} \qquad (1)$$

ただし，γ: 割引率 $(0 \leq \gamma < 1)$，r_t: 時刻 t で得られる報酬とする．利害の衝突状態は，

$$V_{no\_opp} > V_{conflict} \qquad (2)$$

と表現できる．ここで，V_{no_opp}：相手がいない場合の状態評価値，$V_{conflict}$：利害の衝突が起きて互いに譲らない場合の状態評価値とする．さらに，片方のエージェントが譲歩する場合について，$V_{altruistic}$：譲歩するエージェントの状態評価値，$V_{selfish}$：譲歩されたエージェントの状態評価値とする．

まず最初に，譲歩してもエージェントの状態評価値が下がらない場合が考えられる．これは，

$$V_{no\_opp} \geq V_{selfish} = V_{altruistic} > V_{conflict} \qquad (3)$$

と表現でき，弱い衝突と呼ぶ．前節で述べた太田らの研究では，このカテゴリーを扱っている．二つ目は，譲歩したエージェントが得られる報酬は，譲歩しなかったエージェントよりは減少するが，お互いに譲歩しなかった場合よりは多いケースであり，

$$V_{no\_opp} \geq V_{selfish} > V_{altruistic} > V_{conflict} \qquad (4)$$

と表現でき，これを中くらいの衝突と呼ぶ．最後は，譲歩したエージェントが得る報酬は，お互いに譲歩しなかった場合よりも少ないか等しいケースであり，

$$V_{no\_opp} \geq V_{selfish} > V_{conflict} \geq V_{altruistic} \qquad (5)$$

と表現でき，これを強い衝突と呼ぶ．たとえば，最初にゴールしたエージェントしか報酬が得られないような場合は，これに分類される．強い衝突の場合は，エージェントが譲歩することを単純な強化学習によって学習することはできないため，エージェントが報酬を分配する [20] などの方法が必要となる．本稿では，弱い衝突および中くらいの衝突を扱い，そのような衝突における解として” 個性” と” 社会性” を導入する．

4. 強化学習

前述のように，本論文ではマルチエージェント系に個性・社会性を獲得させる手段として強化学習を用いる．強化学習は，与えられる報酬や罰，つまり，強化信号と環境とのインタラクションを通じて，エージェントが自律的に適切な行動

を獲得する手法として有望視されている．したがって，教師あり学習と比較して柔軟な学習が可能であ．本論文で問題としている利害の衝突回避問題に適用することで，同一学習手法で，状況に応じて適応的に個性・社会性が獲得・分化されることが期待される．

強化学習のアルゴリズムは，様々なものが提案されている．ここでは，簡単のため，離散状態，離散行動の問題を扱う．そこで，離散行動の代表的な学習アルゴリズムである Q-learning を適用する．ただし，通常の Q-learning では，行動後の状態における Q 値の最大値を用いて計算するが，ここでは，オンライン化を考慮し，状態 s_t において行動 a_t を取った際に，

$$maxQ(s_t) = (1-\beta)maxQ(s_t) + \beta Q(s_t, a_t) \qquad (6)$$

とQ値の一次遅れ $maxQ$ を計算し，これを最大値の代わりに用いて，

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha(r_t + \gamma maxQ(s_{t+1})) \qquad (7)$$

とQ値を学習した．ただし，α, β: 小さい定数である．これは，つぎのステップで選択された Q 値を使って学習する Sarsa [21] と Q-learning のちょうど中間的な学習アルゴリズムであり，Sarsa と比較して学習が安定し，Q-learning と比較して，オンライン化できるという利点がある．

4.1　ローカルな報酬

マルチエージェント系における，各エージェントへの報酬の与え方は大きく 2 通り考えられる．

- **ローカルな報酬 …** 各エージェントが得た報酬は，それを得たエージェントのみに与えること．
- **グローバルな報酬 …** あるエージェントが得た報酬を，系の中のその他のエージェントにも配分すること．

利害の衝突回避には，個々の利益をある程度犠牲にすることによって，それと引き替えに全体の利益が得られる側面がある．そのため，マルチエージェント強化学習では，たとえば報酬を全エージェントに均等配分するなどのグローバルな報酬を導入し，系としての最適解を得ることが多い．しかし，エージェント数が多くなると，自分の行動と報酬との相関を知ることが難しく，学習が遅くなるという報告もある [18]．現在，グローバルとローカルの報酬を組み合わせるなど，いかにその報酬を配分すべきかという問題に関心が寄せられている [18) 19)] が，明確な結論はいまだ出ていない．

設計者があらかじめ正解を与えずに，エージェントが学習を通して適切な行動を獲得していくという強化学習本来のコンセプトから考えると，このような報酬配分方法を予め設計するという考え方は不自然であると言える．また，2.3 節で述べたように，本研究では，エージェントは自己の利益のみを追求する存在であるという前提を置いた．人間社会の中で実際に強化学習における報酬にあたるものが，予め定められた方法で他者から配分されるような機構があるとは考えにくい．お互いが相手の存在を無視して自己の最適な行動をとろうとすると，結果的にお互いが目的を達成できなくなる．したがって，前述の中くらいの衝突までならば，相手に譲ったほうが，より良い行動となることから，利他的な行動がローカルな報酬だけでも獲得できるのではないかと考える．

以上のような観点から，本研究では，よりエージェントの自律性を重視するという立場をとり，得られる報酬は配分せず，各エージェントはあくまでもローカルな自己の利益のみを追求する存在とする．

5.　シミュレーション実験

5.1　電車乗降問題の設定

Fig.3 のような，電車の乗客が乗降する問題を考える．乗車する客と降車する客がいた場合，相手の存在を無視してどちらも自分の目的をひたすらに達成しようとすると，お互いに進路を妨害する可能性がある．このとき，電車が空いていれば，乗車側と降車側のどちらが先に進んでも大きな混乱は起きず，系のパフォーマンスにもあまり変化がない．しかし，電車が混んでいると，ホームにいる乗車側の客が相手にいったん進路を譲るという行動をとることによって，結果的には互いに一番スムーズな入れ換えが実現できるのは良く知られた事実である．したがって，混んでいれば社会性が，空いていれば個性が出やすいと考えられる．そして，実際に混雑する場所では，降車優先ルールが適用されている．以上のように，個性・社会性が人間の場合どうなるかがわかりやすいため，この問題は個性・社会性の適応的な発現と分化を示す例として適切であると考えた．

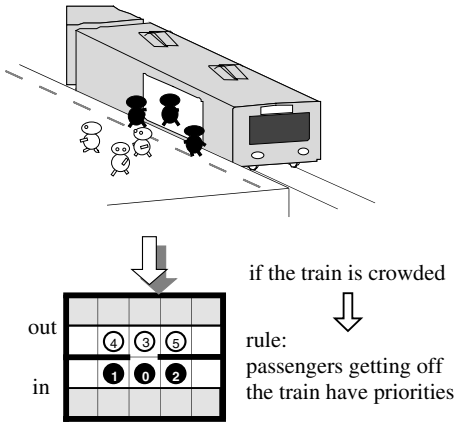


Fig. 3　A sample task in which passengers get on and off a train.

実験環境として，Fig.3 の下の図のような，電車の車内とホームに見立てた格子状の環境を考える．乗車側エージェントと降車側エージェントが向かい合っていて，それぞれ出来るだけ短いステップ数でゴールに到達することが望ましい．ただし，この状況で両者が前に進もうとしても両者ともゴールに到達することはできない．各エージェントは，Fig. 4 のように，自分が乗車か降車か，ドアの位置が前後か左か右

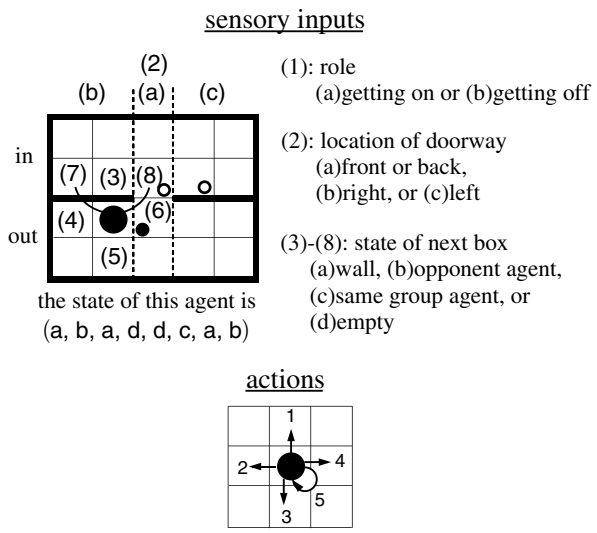


Fig. 4 State and action of each agent.

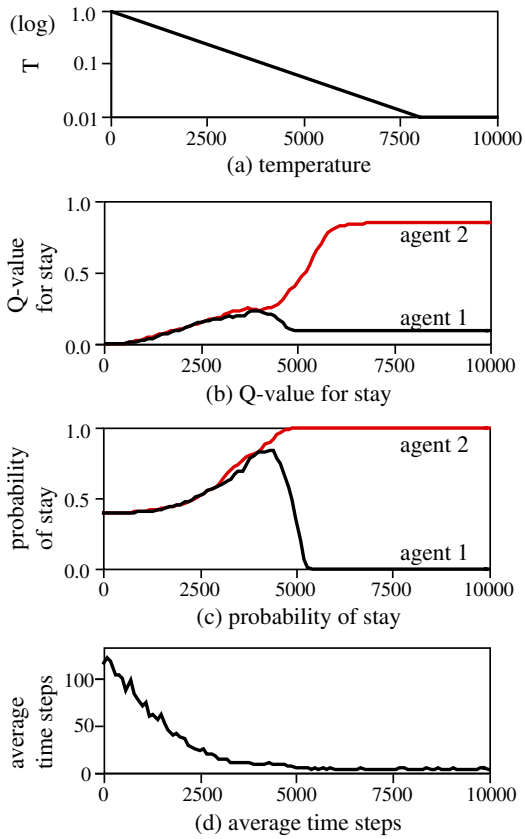


Fig. 5 The change of Q-value at the first time step.

か，また，8 近傍のうち，右後ろと左後ろを除いた 6 箇所に，壁があるか，相手がいるか，味方がいるか，何もいないかの全部で 8 個の信号を得ることができる．ただし，環境を対称とする場合には，自分が乗車か降車かの情報は与えない．行動は，前後左右およびその場に留まるの 5 個の中からボルツマン選択で一つの行動を選択した．ただし，1 つのグリッドには 1 つのエージェントのみが存在可能であり，選択された行動の方向に他エージェントや壁が存在する場合には，移動はしないが，ステップ数はカウントされる．エージェントの初期配置は毎試行ごとにランダムに決定し，行動する順番は，2 エージェントの場合は，毎試行ごとにランダムに，6 エージェントの場合は場所によって決定した．6 エージェントの場合は，以下の図において，番号でその順番を示す．ゴールしたエージェントは，正の報酬 1.0 が与えられ，その後は，右，左，その場にとどまるの 3 つの行動の中からランダムに選択する．すべてのエージェントがゴールするか，またはすべてのエージェントが 200 回行動選択を終了した時点で 1 試行終了とする．Q 値の初期値はすべて 0 とした．

5.2 学習の経過

まず始めに，Fig. 2 で示した 2 エージェントの簡単で全く対称な環境において学習のようすを観察した．学習によって個性が獲得された際の，行動選択に用いたボルツマン選択の温度の変化，真ん中で 2 つのエージェントが向かい合った時に，結果的にその場に留まる行動，つまり，前進とその場に留まる行動に対する Q 値の和の変化のようす，その Q 値に基づいてその行動が選択される確率，および，ゴールまでの平均到達ステップ数を Fig. 5 に示す．ただし，ステップ数は，2 つのエージェントが行動を行なって 1 ステップとする．ボルツマン選択での温度 T は，8000 試行までは

$$log_{10}T = -\frac{Ite}{8000} log_{10}\frac{1.0}{0.01} = -\frac{Ite}{4000} \qquad (8)$$

とし，それ以後は，0.01 で一定とした．ここで，Ite:現在の試行回数である．また，以下のシミュレーションで試行回数を変化させた場合は，上式の 8000 の代わりに試行回数 $\times 0.8$ とした．これらの図より，両者ともいったんは，相手が道を譲ってくれるのを待つため，その場に留まる確率が増えてくるが，その後，片方の Q 値は上昇し，もう片方の Q 値は減少してくる．これは，乱数の影響でたまたま生じた Q 値の差が，温度が小さくなると行動選択に大きく影響するようになり，その差がますますひろがったと考えることができる．

5.3 個性と社会性の分化

5.3.1 エージェントの特定とエージェント数

始めに，前述の 2 エージェント問題について，環境の対称性，および相手を特定できるかどうかが個性・社会性の分化にどう影響するかを調べた．複数のエージェントを用意し，その中から毎試行ごとにランダムに 2 つのエージェントを選んで，Fig. 2 の簡単な環境において学習を行なった．相手が特定できる場合は、自分と相手が同じ位置にいたとしても違う状態と認識し，相手が特定できない場合は同じ状態と認識するものとする．したがって，相手が特定できる場合は，自分以外のエージェントの数に比例して状態空間が大きくなり，特定できない場合は，状態空間の大きさは相手の数とは関係ない．そして，個性と社会性の発現の割合を観察した．Fig. 6 に，(1) エージェントの数を 2, 3, 4，(2) 入力を完全に対称にしたものと，乗車側と降車側とに分けたもの，(3) 相手を特定できる場合とできない場合について，それぞ

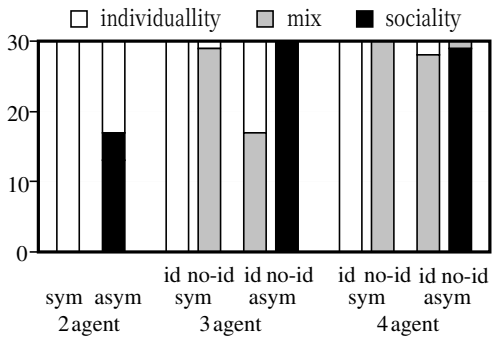


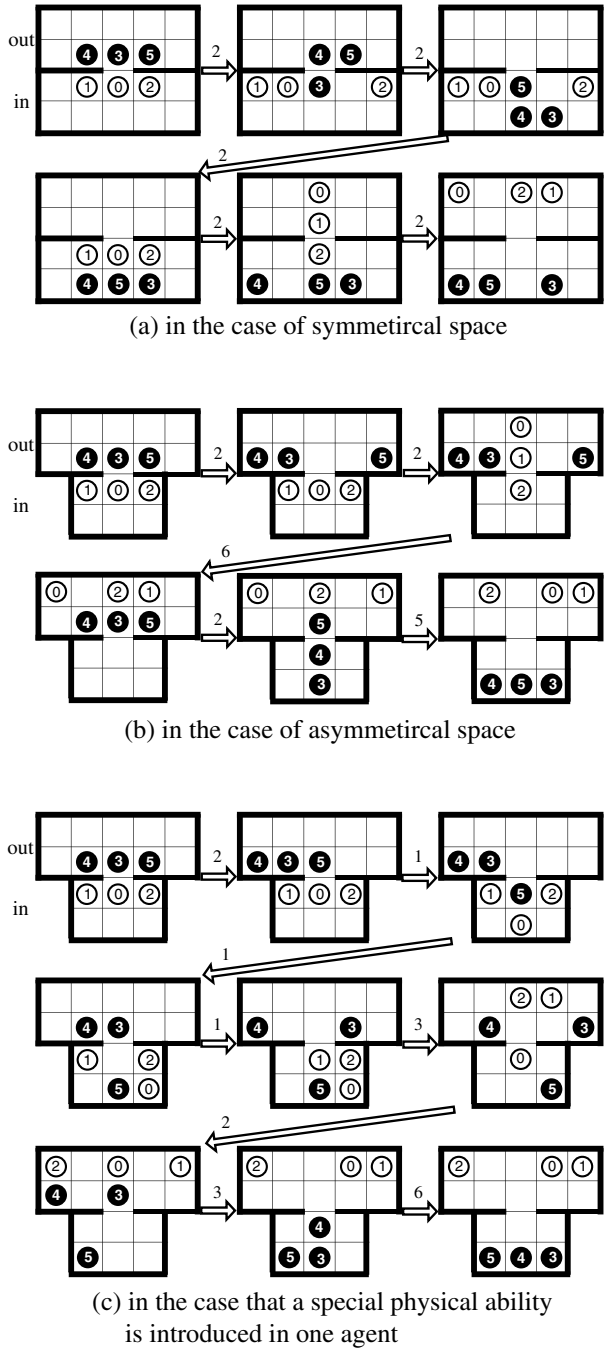
Fig. 6 Differentiation according to the symmetry of environment and identification of the oponent agent in two-agent task.

れ乱数系列を変化させて 30 回学習を行なった時の個性と社会性の割合を示す．ただし，4 エージェントの場合は，一回の学習試行数を 20000 回とした．試行によって，個性になったり社会性になったりする場合もあった．また，3 エージェント以上の場合には，同一試行の中でも，相手によって個性になったり，社会性になったりする場合も観察された．この図より，入力が対称のほうが個性が発現する確率が高く，非対称のほうが社会性が発現する確率が高いことがわかる．さらに，相手が特定できないと社会性が発現しやすく，逆に特定できると個性が発現しやすいことがわかる．

5.3.2　環境の非対称性

入力信号の非対称性による個性と社会性への分化に関しては，前節で 2 エージェント問題の場合に，乗車側か降車側かの信号の入力の有無という形でシミュレーションを行ない，入力が非対称の場合には社会性が，対称な場合には個性が発現しやすいことを示した．本節では，6 エージェントが乗車と降車それぞれ 3 エージェントずつに別れて学習を行ない，環境の対称性を変化させて，乗車側のスペースを狭くすることを混雑と見立て，車内が混雑している時には降車優先のルールができるかどうかを確認した．始めに，環境が対称な場合と非対称な場合で得られた結果の例を Fig. 7(a)(b) に示す．ただし，両者とも，乗車側か降車側かの信号は与えた． この図は，図と図を結ぶ矢印の上の数字分のステップ間隔でエージェントの状態を表示している．対称な環境の場合には，100 回のシミュレーションを行なったところ，先に行動するエージェント，つまり，乗車側のエージェントが必ず先に道を譲った．一方，非対称な環境の場合には，必ず図のように，降車側のエージェントが先に降り，その後に乗車側のエージェントが乗車した．つまり，この場合には，必ず社会性が発現した．

環境の対称性を少しずつ変化させた時の社会性の発現割合を Fig. 8 に示す． 環境は図にあるように，全く対称な環境から徐々に降車側のスペースを狭くし，非対称性を強くした 4 種類の環境を用意した．そして，全く対称な環境に対しては，乗車側か降車側かの信号を入力した場合としなかった場合の両方を取り上げた．また，それぞれに対し，学習回数を 500000 回 (fast) の場合と 1000000 回 (slow) の 2 つの場合についてシミュレーションを行なった．



(a) in the case of symmetircal space

(b) in the case of asymmetircal space

(c) in the case that a special physical ability is introduced in one agent

Fig. 7 Simulation results in the case of 6 agents.

この場合，乗車側か降車側かの信号も入力しない完全に対称な入力の場合には，個性による解決以外に解がないため，すべての場合で個性による解決となった．しかし，相手の情報が得られないため，衝突することも多く，すべてのエージェントがゴールするまでのステップ数はほかに比べて多かった．それ以外の場合には，すべて社会性による解となった．

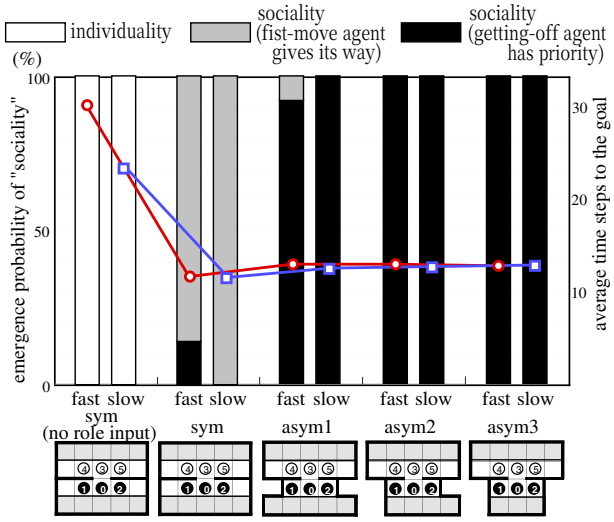


Fig. 8 Differenctiation according to the asymmetry of the environment.

環境が対称で，乗車側か降車側かの区別だけがある場合には，先に行動を行なう乗車側のエージェントが先に乗り込む場合がほとんどであり，ゴールまでのステップ数は最も少なくなった．この場合，環境が対称であるため，先に動くエージェントが道を譲るというルールを獲得したと言える．これは，先に動くエージェントが道を譲った方が，後に動くエージェントが道を譲った場合と比較して，譲られたエージェントがゴールにたどり着くまでのステップ数が少なくなるため，後に動くエージェントが道を譲る確率が小さくなる．したがって，先に動くエージェントが，お互いに道を譲らないで頑張るより，道を譲った方が良いと先に学習するからであると解釈できる．

一方，非対称性が強い場合には，ほとんどの場合，降車側が先に降りるというルールが獲得された．これは，乗車側が先に乗って狭いところにいると，その後降車側の道をふさいでしまう確率が高くなり，降車側のエージェントがゴールにたどりつくまでのステップ数が大きくなってしまう．それで，降車側のエージェントは乗車側のエージェントに比べ，道を譲ることで大きく損をすることになるため，道を譲る行動を選択する確率が減少し，学習によってその差が拡大されたと解釈できる．このように，同じ社会性が獲得された場合でも，獲得されたルールは環境に応じて適応的に変化する能力も持っていることがわかる．

また，学習回数を少なくすると，対称な環境の場合でも降車優先のルールができることがあったり，非対称な環境でも乗車優先となるケースがあった．これは，行動選択の確率に影響する温度が急激に降下することになるため，個々のエージェントが学習しても最適な解に行かない場合が多くなり，結果として多様性が出ると考えられる．

5.3.3 身体特性の差

つぎに，身体特性が違う場合についてシミュレーションを行なった．ここでは，前述の非対称な環境において，一つのエージェントだけ，自分の前にいる相手のエージェントを，その向こう側が空いている場合に押しのけて前に進むことができるという設定にして学習を行なった．そのときの学習後の行動の例を Fig.7(c) に示す．身体特性が全く対称な場合には，Fig.7(b) のように，エージェントの配置によらず，降車側優先の社会性が発現している．しかし，力が強いエージェントは，乗車側でかつドアの前にいないときでも，ほかの乗車側エージェントが道を譲っているにもかかわらず，ドアの前に進み，降車側エージェントよりも先に乗車し，その後，降車側エージェントが降車し，それから残りの乗車側エージェントが乗車した．エージェントの配置を変化させた場合にも，力が強いエージェントが常に他のエージェントを押しのけて最初に乗車または降車した．つまり，この場合は，降車優先の社会性が破られ，個性が発現したと言える．もちろん，他のエージェント同士では社会性が見られた．

6. まとめ

マルチエージェント環境において，利害の衝突回避のために個性と社会性が発現するという考え方を提唱した．そして，単に個々のエージェントに自らが獲得した報酬に基づいて強化学習を行なうことによって，個性や社会性が発現することを確認した．また，個性と社会性に分化する要因として，環境の非対称性，エージェントの数と特定，および個体間の物理的能力の差をあげた．そして，電車の乗り降りのシミュレーションを通して，非対称環境において，降車側優先のルール（社会性）の形成を確認するとともに，上記の分化要因によって実際に個性と社会性に適応的に分化することを確認した．また，その分化の仕方は，われわれの人間社会での分化からみて合理的なものであると考えられる．

謝辞

本研究の一部は，文部省 (文科省) 科学研究費重点領域研究「創発システム」(No. 264)，基盤研究 B(No. 10450165, No. 14350227)，若手研究 B(No. 13780295) および学術振興会未来開拓学術研究推進プロジェクト「生物的適応システム」(JSPS-RFTF96100105) の補助の下で行われました．ここに謝意を表します．

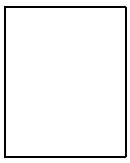
参 考 文 献

1）石田亨, 片桐恭弘, 桑原和宏: 分散人工知能, コロナ社 (1996)
2）J. メイナード-スミス (寺本英, 梯正之訳): 進化とゲーム理論 -闘争の論理-, 産業図書株式会社 (1985) Maynard Smith J: Evolution and the Theory of Games, Cambridge Univ. Press (1982)
3）市川純章, 原文雄: 性格をもったロボットの群行動シミュレーション, 第 8 回自律分散システムシンポジウム資料, 349/354, 計測自動制御学会 (1996)
4）Muraciano, A., Millan, J.R. and Zamora, J.: Specialization in Multi-agent Systems through Learning, Biological Cybernetics, **76**, 375/382 (1997)
5）Zamora, J., Millan, J.R. and Murciano, A.: Learning and stabilization of altruistic behaviors in multi-agent sys-

tems by reciprocity, Biological Cybernetics, **78**-3, 197/205 (1998)
6）デュルケーム, ジンメル: 「社会的分化論」世界の名著, 中央公論社 (1968)
7）E.P. オダム: 生態学の基礎 上巻, 三島次郎訳, 培風館 (1974)
8）ホッブス (水田洋, 田中浩訳): 「リヴァイアサン」世界の大思想, 河出書房新社 (1974)
9）寺邊正大，椹木哲夫，片井修，鷲尾隆: マルチエージェント環境下での情報の共有と組織学習, 第 5 回マルチエージェントと協調計算ワークショップ (MACC), 講演配布資料 (1995)
10）太田順, 横川洋一, 新井民夫: 複数自律ロボット系における漸進的戦略形成に関する研究, 日本ロボット学会誌, **14**-3, 379/385 (1996)
11）柴田 克成, 伊藤 宏司: 認識の学習に基づく注意と連想記憶の形成, 電子情報通信学会技術研究報告, NC99-137, 153/160 (2000)
12）柴田克成, 伊藤宏司: 利害の衝突回避のための交渉コミュニケーションの学習と個性の発現 -リカレントニューラルネットを用いたダイナミックコミュニケーションの学習-, 計測自動制御学会論文誌, **35**-11, 1346/1354 (1999)
13）柴田克成: 強化学習とロボットの知能 ―あめとむちで知能は作れるか？―, 2002 年人工知能学会全国大会 パネルディスカッション「強化学習の諸相とその展望」原稿 (2002)
14）柴田克成, 岡部洋一, 伊藤宏司: ニューラルネットを用いた強化学習-センサからモータまでの合目的的・調和的学習-, 計測自動制御学会情報・システム部門学術講演会 2001, 227/232 (2001)
15）柴田克成, 岡部洋一, 伊藤宏司: ニューラルネットワークを用いた Direct-Vision-Based 強化学習, 計測自動制御学会論文集, **37**-2, 168/177 (2001)
16）アダム・スミス (大内兵衛, 松川七郎訳): 諸国民の富, 岩波書店 (1969)
17）リチャード・ドーキンス (日高敏隆他訳): 利己的な遺伝子, 紀伊国屋書店 (1991)
18）白川英隆, 木村元, 小林重信: 強化学習による協調的行動の創発に関する実験的考察, 第 25 回知能システムシンポジウム資料, 119/124 (1998)
19）宮崎和光, 石原秀一, 荒井幸代, 小林重信: 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, 第 11 回自律分散システムシンポジウム資料, 289/294 (1999)
20）Katsunari Shibata & Koji Ito: Autonomous Learning of Reward Distribution for Each Agent in Multi-Agent Reinforcement Learning, Intelligent Autonomous Systems, **6**, 495/502 (2000)
21）Richard S. Sutton & Andrew G. Barto: Reinforcement Learning, MIT Press (1998)

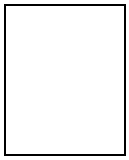
［著　者　紹　介］

柴　田　克　成（正会員）



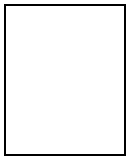
shibata@cc.oita-u.ac.jp, 1989 年 東大大学院工学系研究科機械工学専攻修士課程修了．1989 年 (株) 日立製作所 に入社．1992 年 10 月 同社退職．1993 年 東大大学院工学系研究科先端学際工学専攻博士課程中退．1993 年 東大先端科学技術研究センター助手．1997 年 東工大大学院総合理工学研究科リサーチアソシエイト（日本学術振興会未来開拓学術研究推進プロジェクト研究員）．2000 年 大分大学工学部電気電子工学科講師．2001 年 同助教授．主として，ニューラルネットを用いた強化学習・自律学習システムの研究に従事（博士（工学））．

上　田　雅　英



ueda@ito.dis.titech.ac.jp, 1999 年 東京工業大学大学院総合理工学研究科知能システム科学専攻修士課程修了．いすゞ自動車 (株) を経て，2001 年 (株) 本田技術研究所栃木研究所勤務．

伊　藤　宏　司（正会員）



ito@dis.titech.ac.jp, 1969 年 名大大学院工学系研究科応用物理工学専攻修士課程修了．1970 年 同大工学部自動制御研究施設助手．1979 年 広島大学工学部電気系助教授．1992 年 豊橋科学技術大学情報工学系教授．1996 年 東工大大学院総合理工学研究科教授．主として，運動制御，ロボティクス，マンマシンインターフェースの研究に従事（工学博士）．