[計測自動制御学会論文集
Vol.35, No.11, 1/9（1999）]

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†

－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－

柴 田 克 成*・伊 藤 宏 司*

Learning of Communication for Negotiation to Avoid Some Conflicts of Interests

- Learning of Dynamic Communication Using a Recurrent Neural Network-

Katsunari SHIBATA* and Koji ITO*

We believe that communication in multi-agent system has two major meanings. One of them is to transmit one agent's observed information to the other. The other meaning is to transmit what an agent is intending. In such communication, dynamic communication with a communication loop is required. Here we focus the latter, and aim to the emergence of the autonomous and decentralized arbitration through communication among some agents. The communication contents and representation of them are not prescribed and are acquired by learning using a reinforcement signal, which is given to the agent after its action. The reinforcement signal is not shared with the other agents. Since the agent often has to make a decision from the past communication signals, the architecture using recurrent type (Elman) neural network is proposed. The ability of this architecture was examined by two and four agent negotiation problems. A variety of negotiation strategies (individuality) to avoid the conflicts after their decisions emerged among them through the simple learning with the simple architecture.

Key Words: dynamic communication, negotiation, reinforcement learning, recurrent neural network, individuality, multi-agent system

1. はじめに

コミュニケーションはマルチエージェントシステムにおける協調や調停に有効である．しかし，与えられたシステムに対し，何をコミュニケーションし，それをどのように表現し，どのような順序でコミュニケーションすれば良いかを設計することは一般に難しい．また，それを予め設計してしまうことで，環境の変化に対応できず，柔軟性を失ってしまうことも考えられる．われわれ人間は，非常に複雑なコミュニケーションを行うことができる．しかし，環境によって異なった言語を獲得することを考えると，この能力は，予め遺伝情報などの形ですべて与えられるのではなく，個々の学習によって獲得される部分が大きいと考えられる．したがって，マルチエージェントシステムにおいて創発的手法によるコミュニケーションの獲得はたいへん重要な課題であるとともに，生物をヒントとして，実現可能ではないかと考えられる．

筆者らは，コミュニケーションには Fig. 1 に示すように，大きく2つの意味があると考えている．一つは，**エージェントの観測結果を他のエージェントに伝え**，受け手の観測の不十分さを補うことである．Werner らの研究[1]では，目を持つが動けないメスが信号を発し，目が見えないオスがそれを

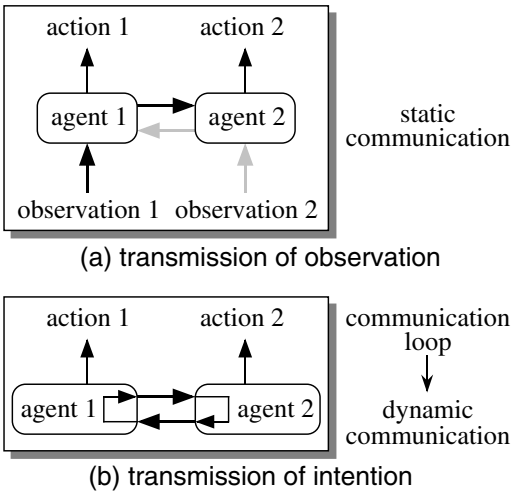


Fig. 1 Two meanings of communication. Transmission of observation can be static, and transmission of intension has to be often dynamic because it needs a communication loop for negotiation between agents.

† 第 11 回自律分散システムシンポジウムにて発表（1999・1）

* 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，横浜市緑区長津田町 4259

* Dept. of Comp. Intel. and Sys. Sci., Interdisciplinary Graduate School of Sci. and Eng., Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama

（Received January 7, 1999）
（Revised June 14, 1999）

受信して行動する．そして，出会ったペアのみが子孫を残すことによって，出会うために有効なメスからオスへのコミュニケーションを進化的に獲得している．一方，Nakanoらが行った共通言語の創発に関する先駆的な研究[2]では，複数のロボットが同時に同じ対象を観察する際に，学習によってその対象を表す共通言語を獲得させている．これによって，ロボットが対象を見ていなくても，対象を見ているもう一方のロボットからの言葉によって対象の存在を知覚できる．これも，観測結果の伝達が目的であると言える．

コミュニケーションのもう一つの意味は，**自分の意志を伝えること**である．これは，複数のエージェント間の協調行動や利害の衝突を避ける際に有効である．このようなコミュニケーションに関する研究は，すでに分散人工知能の分野で行われている[3)4)]．しかし，いずれの場合も，コミュニケーションの内容やその表現方法および戦略が予め設計され，その設計の善し悪しが議論されている．

この2つのコミュニケーションの性質をまとめたものをFig. 1に示す．観測結果の伝達では，エージェントが観測したことを相手に伝える．したがって，基本的には，情報の流れは一方通行である．しかし，意思の伝達では，利害の衝突が起きた場合にそれを調整するために交渉しなければならないため，**コミュニケーションのループ**が必要となる場合が多い．つまり，**動的なコミュニケーション (Dynamic Communication)**が必要となる．したがって，その内容や表現だけでなく，いかにコミュニケーションを進めるかという戦略が重要となる．また，その際，過去のコミュニケーションの経緯を反映させる必要性が生じることも予想される．

以上より，本研究では，「**コミュニケーションの学習による獲得**」と「**自分の意志の伝達**」に焦点を当てる．まず，予めコミュニケーションの内容や戦略を与えなくても，適切なコミュニケーション信号と行動を生成するためのエージェントの構成と，行動後に受け取る報酬や罰（強化信号）から学習する方法を提案する．ただし，学習によって獲得されたコミュニケーションが，エージェントの意思であるかどうかを判定することは困難である．そこで，ここでは，各エージェントへの入力を過去のコミュニケーション信号のみとし，学習に用いる外部からの信号を，最終的な行動時の衝突の情報のみとすることにより，本論文中でのコミュニケーションを行動のための「意思」と解釈した．また，過去の経緯を反映させたコミュニケーションを実現するため，リカレント型のニューラルネットを用いて出力を決定する．そして，エージェントが常に報酬を得るためには，相手のエージェントに応じて行動を変化させなければならないエージェント間交渉問題を取り上げて，そのコミュニケーション能力を検証する．このようなエージェント間の利害の衝突を回避するためには，エージェントによる行動の違い（個性），環境による行動の違い（社会性）を利用する方法があることが示されている[5)]．ここでは，交渉するエージェント間の環境を対称なものと設定したため，**エージェント間で交渉の戦略に個性の発現**が必要である．そこで，学習前の入出力関係がすべて同じエージェントに対し，同一の学習則を適用して，様々な個性が発現するかを観察する．また，マルチエージェントシステムにおける強化学習では，強化信号のエージェント間配分が問題となる[6)7)]．しかし，外部の設計者が予め強化信号の配分を決定することは，エージェントが自律分散的に行動し学習する本研究の意図に反する．さらに，われわれ人間や生物の社会でも予め決められた方法で報酬や罰の配分は行われているとは考えられない．したがって，ここでは，各エージェントが獲得した報酬や罰は他のエージェントと配分しない．そして，配分することなくすべてのエージェントが満足する解を得られるかどうかも本研究の一つのポイントである．

2. エージェントの構成と学習方法

2.1 構成

Fig. 2に，提案するエージェントの構成を示す．エージェントは，他のエージェントおよび自分の1単位時間前のコミュニケーション信号を受け取り，ニューラルネットに入力する．ニューラルネットは，Elman型のリカレントネット[8)]を用いた．つまり，1単位時間前の中間層の値も入力値として与えられる．これにより，過去の履歴を利用した交渉や行動の決定が可能になる．ニューラルネットの各ニューロンの出力関数は，-0.5から0.5の値域を持つシグモイド関数とした．出力は，コミュニケーション用と行動用の2つ用意した．コミュニケーション信号と行動は1と-1のどちらかとし，1を出す確率はニューラルネットの対応する出力に0.5を足した値とした．エージェントは，コミュニケーション信号を同期して交換し合い，最後に行動を決定する．ただし，コミュニケーション時の行動出力および行動決定時のコミュニケーション出力は用いず，学習も行わない．交渉の最初は中間層からのフィードバック信号も含めて入力をすべて0とした．これによって，最初のコミュニケーション信号が1である確率は，相手のエージェントによらず常に一定となる．

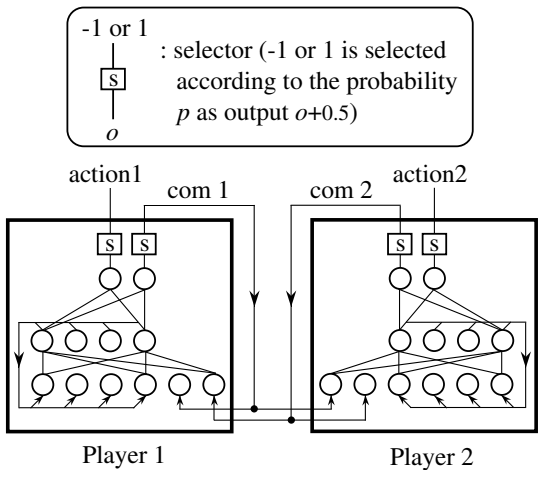


Fig. 2 Proposed architecture of an agent for communication using Elman type recurrent neural network.

2.2　学習方法

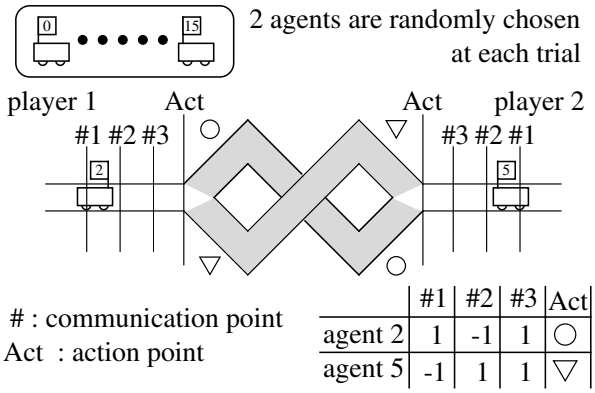
エージェントは，行動後に得られる強化信号をもとにコミュニケーションと行動を学習する．エージェントが報酬を得たときには，その際にとっていたコミュニケーションと行動の生起確率を大きくし，罰の場合には小さくする．ニューラルネットは BP(Back Propagation)[9] または BPTT(Back Propagation Through Time)[9] という教師あり学習で学習した．3 回のコミュニケーション出力と 1 回の行動出力のそれぞれに対してつぎのような教師信号を内部生成してネット内の重み値とバイアス値を共に学習する．

$$x_{ideal} = x + \eta \cdot r \cdot x' \cdot o \qquad (1)$$

ただし，x_{ideal}：教師信号，$x = f(u) = 1/(1+exp(-u))-0.5$：ニューラルネットの出力，$u$：出力ニューロンの内部状態，$\eta$：学習係数 (ここでは 0.1)，$r$：強化信号 (ここでは報酬の場合 1，罰の場合 -5)，$x' = dx/du = (0.5-x)(0.5+x)$，$o$：実際のコミュニケーション信号または行動（ニューラルネットの出力ではなく，確率的に決定した後の値）．x' は，確率 p が 0 か 1 に近い場合に，安定になるように加えた．また，ここでは，簡単のため，通常の強化学習とは異なり，コミュニケーション出力の学習は学習終了後に時間をさかのぼって行い，同一試行中の教師信号の生成にはすべて同一の強化信号 rを用いた．入力層から中間層への結合の初期値は微小な乱数とし，中間層から出力層への結合および中間層と出力層のバイアス値はすべて 0.0 とした．これによって，入力層と中間層の間の結合の初期重み値は各エージェント毎に違うものの，学習前の出力の値は，全エージェントとも入力によらず，すべて 0.5 となる．教師信号は，コミュニケーションを行った各時刻でのコミュニケーション出力および最後の行動出力のそれぞれについて与えられる．BPTT では，各教師信号が与えられた時間から交渉の始めまで，誤差信号が入力層からフィードバック結合を通って中間層に戻ることによって時間をさかのぼって学習する．一方，通常の BP では，誤差信号はフィードバック結合を通って時間をさかのぼらず，教師信号を与えた時間での状態のみに基づいて重み値が更新される．これらの学習を通して，各エージェントの入出力関係は，当初は全く同じであったものが，徐々に変化し，各エージェント毎に異なる入出力関係が実現される．これを，ここでは，個性と呼ぶ．この個性は，各エージェントのニューラルネットの入力層から中間層の間の結合の初期重み値，ニューラルネットの出力からコミュニケーション信号，行動を決定する際の確率的要因およびエージェントがタスクに選ばれる確率的要因に由来し，学習によって適切な個性へと発展する．

2.3　バッファ方式との比較

ダイナミックコミュニケーションの学習を行うためには，過去の相手と自分のコミュニケーション信号を保持し，過去から現在までのコミュニケーション信号を入力として与え，リカレント機構のない通常の階層型ニューラルネットで学習させる方法も考えられる．これを，バッファ（シフトレジスタ）方式と呼ぶ．この場合，過去何回分のコミュニケーション信号を保持し，入力として与えるかを予め決める必要がある．さらに，コミュニケーション回数が増大すると，それだけ入力数を増やさなければならないため，ニューラルネットの規模が大きくなり，メモリ量，計算量の増大が考えられる．ただし，リカレントニューラルネットを利用した場合と比較して，複雑な交渉パターンの学習が容易に行われることが期待される．しかし，リカレントニューラルネットを用いた場合は，中間層の内部状態がつぎの時間に入力としてフィードバックされるため，解（報酬が得られる経路選択）という固定点に収束するダイナミクスが学習によって獲得される可能性が高いと考えられる．



	#1	#2	#3	Act
agent 2	1	-1	1	○
agent 5	-1	1	1	▽

Fig. 3　Two-agent negotiation problem.

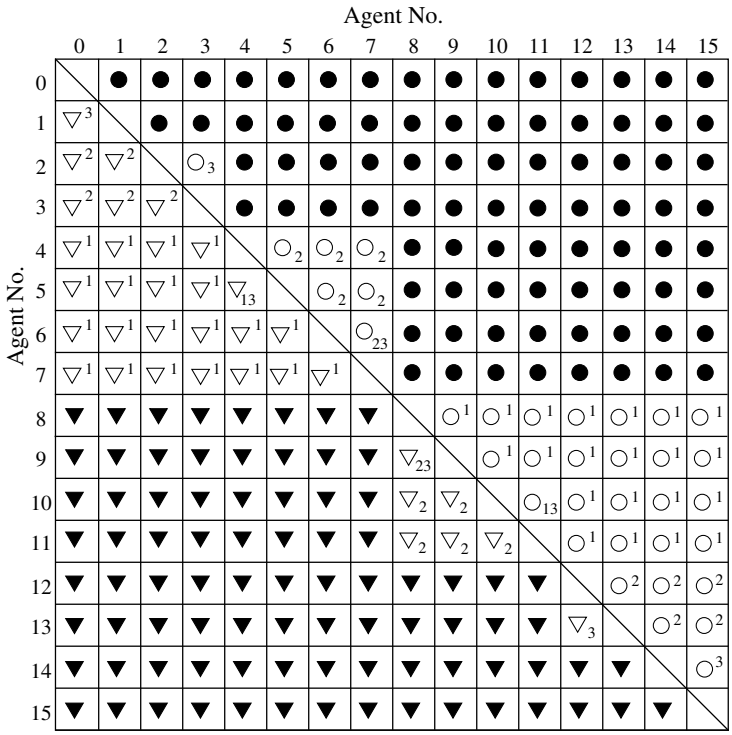
3.　シミュレーション

3.1　2エージェント交渉問題

3.1.1　設定

始めに，2エージェント交渉問題について述べる．Fig. 3 にシミュレーションの環境を示す．2エージェント交渉問題では，まず，予め用意した 16 エージェントの中から 2 つのエージェントをランダムに選ぶ．エージェントは，相手がどのエージェントかを知らされることなく，3 回のコミュニケーションの機会＃1，＃2，＃3が与えられ，各機会ごとに，前節のように，1か－1のコミュニケーション信号を出力する．最初の機会には，相手のコミュニケーション信号を知ることなく出力を決定する．2回目は，自分と相手の1回目のコミュニケーション信号を入力として出力を決定する．3回目には，Elman ネットが 1 回目のコミュニケーション信号に関する情報を保持していれば，1回目と2回目の両方の信号を考慮して出力を決定することができる．そして最後に，エージェントは自分が通るルートを○ (行動出力=1) か▽ (行動出力=-1) のどちらかに決定する．もし，両者が同じルートを選んだ場合は罰 ($r = -5$) が，違うルートを選んだ場合は報酬 ($r = 1$) が与えられる．報酬より罰の絶対値を大きくしたのは，1 回でもエージェントが衝突した場合は，コミュニケーションや行動の戦略を変化させ，系を不安定にし，新たな解を探索させるためである．罰の絶対値が小さすぎる

Table 1 A series of 3 communication signals and action for each agent pair after learning, and some examples.



example

(1)

agent	#1	#2	#3	Act
0	1	1	1	○
1	1	1	1	▽

(2)

agent	#1	#2	#3	Act
0	1	1	1	○
2	1	1	-1	▽

(3)

agent	#1	#2	#3	Act
1	1	1	1	○
2	1	1	-1	▽

(4)

agent	#1	#2	#3	Act
0	1	1	1	○
8	-1	-1	-1	▽

(5)

agent	#1	#2	#3	Act
4	1	-1	1	○
5	1	-1	-1	▽

(6)

agent	#1	#2	#3	Act
6	1	-1	-1	○
7	1	-1	-1	▽

(7)

agent	#1	#2	#3	Act
12	-1	-1	1	○
13	-1	-1	1	▽

と，エージェントが衝突する場合が多少あっても安定してしまい，解が得られないことが多かった．また，逆に罰の絶対値が大き過ぎると，系が非常に不安定になり，収束しなかった．この強化信号をもとに，エージェントは前節の要領で３回のコミュニケーション信号と行動を学習する．この問題では，コミュニケーションの機会が３回なので，そのパターンから最大 $2^3 = 8$ のエージェントの区別が可能である．さらに，コミュニケーションパターンが同じ２つのエージェントが違う行動をとることもできるので，理想的には $2^4 = 16$ のエージェントの中から任意の２つのエージェントを選んでも常に違うルートをとることが可能である．

3.1.2　結果

Table 1 の右側に，学習が成功したときの交渉の例を示す．学習が成功したほとんどの場合，各コミュニケーション信号や行動の選択確率は０か１に近い値になり，確率的要因はほとんどなくなった．最初の３つの例について詳しく説明する．

(**例１**) エージェント０とエージェント１が選ばれた場合．両者は３回ともコミュニケーション信号１を出力し，最終的にエージェント０はルート○を，エージェント１はルート▽を選択した．

(**例２**) エージェント０とエージェント２が選ばれた場合．エージェント０は例１と同一のコミュニケーション信号と行動をとった．一方，エージェント２は，コミュニケーション信号１を２回出力した後，３回目は－１を出力．そして，ルート▽を選択した．

(**例３**) エージェント１とエージェント２が選ばれた場合．エージェント１は，コミュニケーション信号１を３回出力した後，ルート○を選択．一方，エージェント２は例２と同様のコミュニケーション信号と行動をとった．

コミュニケーション信号の意味は学習前にエージェントに与えなかったが，学習後に獲得されたコミュニケーション信号に筆者らが意味づけを行った．上述の例におけるエージェント０のように，相手のエージェントによらず常に同じルートをとったエージェントは，そのコミュニケーション信号のシーケンスも相手によらず常に一定であった．このようなエージェントを，ルート○を通ることを主張し続けたと定義する．したがって，例１のエージェント１は，コミュニケーションの３回の機会でルート○を通ることを主張し続けたが，最終的に，エージェント０に譲歩し，ルート▽を選択したと解釈できる．ところが，例３では，相手のエージェント２が先に譲歩したため，主張どおりルート○をとったと解釈できる．

Table 1 に，すべてのエージェントペア間での交渉結果を示す．表に出てくるエージェントの順番は，ルート○をとる確率が多いものから順に並べた．表中の●は，上述の例におけるエージェント０のように，ルート○をとることを主張し続け，最終的に主張どおりルート○をとったことを示す．上付き添え字のある○は，例７におけるエージェント 12 のように，添え字の番号のコミュニケーションの機会までルート▽を主張したが，そのつぎの機会に主張を変更して最終的にルート○をとったことを示す．下付き添え字のある○は，例５のエージェント 4 や例６のエージェント 6 のように，最初はルート○を主張したものの，添え字の番号のコミュニケー

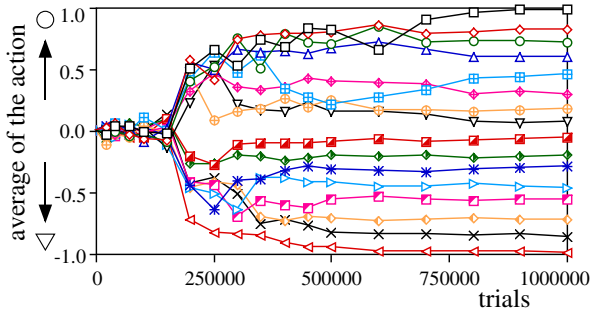


Fig. 4　The change of the action probability of each agent according to the progress of learning.

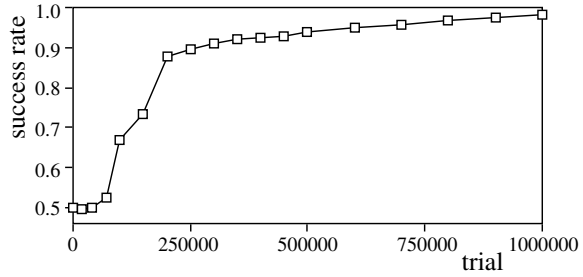


Fig. 5　The change of the success rate.

ション機会の際にルート▽を主張し，さらにその後は再びルート○を選択したことを示す．▼と▽は，最終的にとったルートが▽の場合であるが，見方はそれぞれ●と○と同様である．この表から，各エージェントは，自律分散的に集団の中で順序づけされ，その順序が上のものは○を下のものは▽のルートを通ることによって衝突を避けていることがわかる．また，エージェント 0 から 7 までは，最初のコミュニケーション出力が 1 で，エージェント 8 から 15 までは-1 である．そして，選ばれた 2 つのエージェントの最初のコミュニケーション信号が異なっていると，つまり，Table 1 の右上および左下の部分では，コミュニケーション信号は最後まで変化せず，最初のコミュニケーション信号が 1 の場合は○を 0 の場合は▽のルートを選択していることがわかる．それに対し，Table 1 の左上および右下の部分では，最初のコミュニケーション信号が同じ，つまり，両者の主張に衝突があり，エージェント 1 または 15 に近いほど頑固で主張をなかなか変えず，エージェント 7 または 8 に近いほど，自分の主張をすぐに変えていることがわかる．

Fig. 4 に，各エージェントの最終的な行動出力（ルート）の平均値が学習によってどのように変化しているかを示す．この図から，学習の初期では，好みの経路が○であるか▽であるかの 2 つの集団に大きく別れ，その後，細かく順序付けされ，その際，エージェント間で順序が入れ替わるようなこともあることがわかる．また，学習がうまくいかなかった場合のいくつかを見ると，学習の初期の 2 分化の際に，8 エージェントづつに分かれず，7 エージェントと 9 エージェントのようになり，その配分が学習終了まで変化していなかった．つまり，最初の 2 分化は，各エージェントにおいて，後の学習で変更できないような変化が起こっているものと考えられる．また，Fig. 5に，学習が成功した場合の成功率の変化を示す．2 分化される少し前から成功率は大きく上がり始め，2 分化の完了後は徐々に上がっていることがわかる．

また，ここで示したシミュレーションでは，ルート○の主張は，(1, 1, 1) であったが，シミュレーションによっては，(-1, -1, -1), (1, -1, 1), (-1, 1, -1) のこともあった．これは，各エージェントのニューラルネットの初期値の値，および，タスク中での確率的要因のためである．しかし，(1, 1, -1), (1, -1, -1), (-1, 1, 1), (-1, -1, 1) にはならなかった．これは，自分の前回のコミュニケーション信号も入力となるため，常に前の信号を繰り返すか，常に前の信号の反対の信号を出すことを学習することがニューラルネットにとって容易であるからと考えられる．また，エージェントの数を 8 に減らすと，解はより簡単に得られた．この場合，2 回以上主張を変えるエージェント，つまり，Table 1 で下付き添え字のあるエージェントは出現しなかった．このことから，2 回以上主張を変えることは学習が難しいと考えられる．

この問題は，各エージェントの中間層ニューロン数 2 個だけで解くことができた．ただしその場合，解に至る確率は非常に小さかった．このとき，複雑なコミュニケーションを行っている Table 1 のエージェント 2 を例として，シンプルなリカレントニューラルネットでどのようにしてコミュニケーションを行っているかを示す．Fig. 6 にニューラルネットの結合の重み値およびニューロンのバイアス値を示す．そして，相手のエージェントを変えた 4 つの場合について，コミュニケーションを進めるにしたがって中間層の値と出力（コミュニケーション時にはコミュニケーション出力，行動決定時には行動出力）の値が変化するようすを Fig. 7 に示す．これらから，このエージェントの基本的な戦略は，

1. 中間層から両出力への結合の重み値はほぼ等しいので，両出力は常にほぼ同じ値となる．
2. 中間層ニューロン 2 のバイアス値が正で，そこからコミュニケーション出力への結合の重み値が正なので，交渉の最初のコミュニケーション出力は 1 となる．
3. 相手のコミュニケーション信号が －1 ならば，入力層ニューロン 1 →中間層ニューロン 2 の負の結合と中間層ニューロン 2 から出力層への正の結合によって，両出力は 1 になる．これは，Fig. 7 の (b)(c)(d) でも観察できる．
4. 入力層ニューロン 2 →中間層ニューロン 1，中間層ニューロン 1 →出力ニューロン 1 の 2 つの負の結合と，入力層ニューロン 4 →中間層ニューロン 2 への正のフィードバック結合により，前の出力を保持しようとする．
5. 相手のコミュニケーション信号が 1 のとき，入力層ニューロン 1 →中間層ニューロン 2 への負の結合により中間層ニューロン 2 の出力が減少する．そして，自分と相手のコミュニケーション信号が共に 1 の場合が 2 回続くと，中間

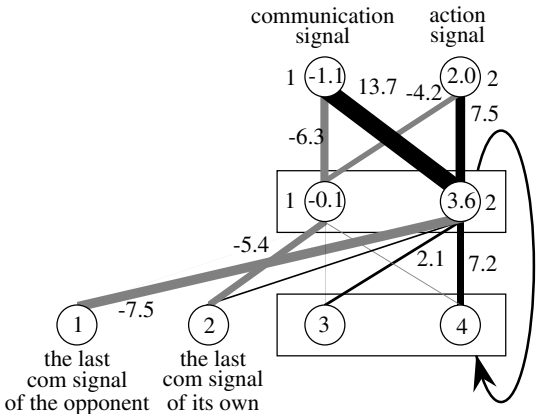


Fig. 6 Connection weights of the neural network in the agent 2 after 2-agent negotiation problem.

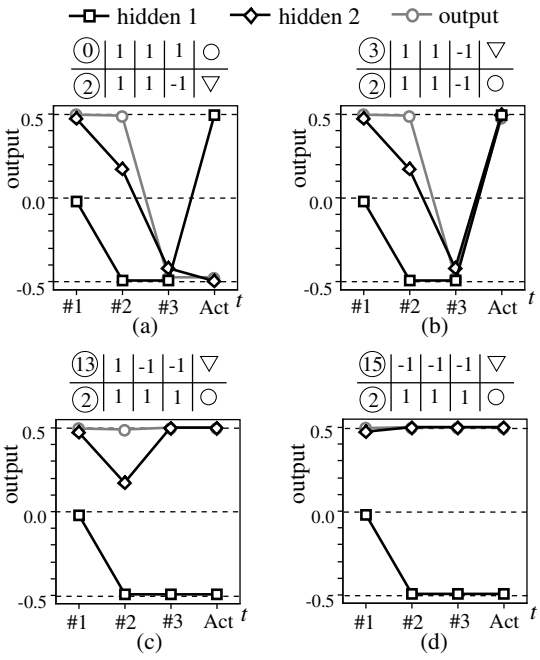


Fig. 7 Time series of the hidden neurons' outputs of the agent 2 depending on the opponent agent.

層ニューロン２の出力がさらに下がって０より小さくなり，コミュニケーション出力が－１となる．これは，Fig. 7 の (a)(b) で観察される．

このニューラルネットは，中間層が過去のコミュニケーションの履歴の特定のパターンを表現しているわけではない．したがって，過去のコミュニケーション信号から出力へのテーブルを単に学習したのではなく，解に収束するダイナミクスを学習したと解釈できる．また，項目５では，コミュニケーションの２回目と３回目の機会で入力が全く同じであるにもかかわらず，出力が変化している．これは，中間層から入力層へのフィードバック結合で過去の履歴が保持されたことによるものと言える．リカレントでない通常の階層型ニューラルネットで学習させると，中間層ニューロン数を増やしても，このような機能を持つエージェント，つまり，Table 1 におけるエージェント２，３，１２，１３は出現しなかった．

ニューラルネットの中間層ニューロン数を４個とし，結合の初期値を変化させて 100 回のシミュレーションを行った．120 すべての組合わせで衝突がなかったのは，BPTT で学習させた場合 41 回で，通常の BP の場合 52 回だった．Elman 型ニューラルネットでは，中間層からフィードバックされた入力と中間層との結合は，文脈情報（過去の履歴の中で現在の状態に影響を与える情報）の保持と出力への反映の両方に使われる．BPTT では文脈の保持まで考慮した学習ができるが，通常の BP では文脈情報の出力への反映だけが学習される．しかしこの場合は，文脈情報を出力へ反映させる学習が，文脈の保持にも有効であったと考えられる．

2. 3で述べたバッファ方式でも同様のシミュレーションを行った．バッファ方式では，相手と自分の過去 3 回分のコミュニケーション信号を入力とするため，入力の数は 6 個となり，中間層ニューロン数は 4 個とした．そして，試行開始時のバッファ(シフトレジスタ) の初期値はすべて 0 とした．この場合，ニューラルネット内のパラメータ（結合の重み値，バイアス値）の数は，リカレントネットの場合と同じである．この結果，53 回で解が得られ，リカレントネットの場合と比較して，大差がないことがわかった．

3. 1. 3　収束ダイナミクスの検証シミュレーション

この学習によって，解という固定点に収束するダイナミクスが学習されていれば，冗長な回数のコミュニケーション機会で学習を行っても，より少ないコミュニケーション機会で交渉が成立する可能性が高いと考えられる．そこで，前述のシミュレーションを，5 回のコミュニケーション機会で学習させた後に，与えるコミュニケーション機会を 3 回に減らしても利害の衝突が回避されるかどうかを検証した．この場合，バッファ方式の入力数は $5 \times 2 = 10$ 個となる．その結果，いずれの場合も，5 回のコミュニケーション機会の後に 120 のエージェントの組合わせすべてで衝突を回避することができた．3 回のコミュニケーション機会の後に，衝突が回避されなかった組合わせの数を 100 回のシミュレーションで平均すると，120 の組合わせのうち，リカレントネット (BPTT) の場合は 3.5，標準偏差が 1.5 で，バッファ方式の場合は 6.8，標準偏差が 2.8 となり，いずれの場合も，かなりの確率で衝突が回避されることがわかった．特に，リカレントネットのほうがバッファ方式の約半分の衝突確率であった．これは，前述のように，解という安定点に収束するというダイナミクスが学習によって得られやすく，収束が速くなるのに対し，バッファ方式では，5 回のコミュニケーション機会の後半で，重点的に交渉することを学習することも比較的多かったためと考えられる．

3. 2　４エージェント交渉問題

3. 2. 1　設定

つぎに，４エージェント交渉問題について述べる．前述の２エージェント問題では，全エージェントを順序付けし，選

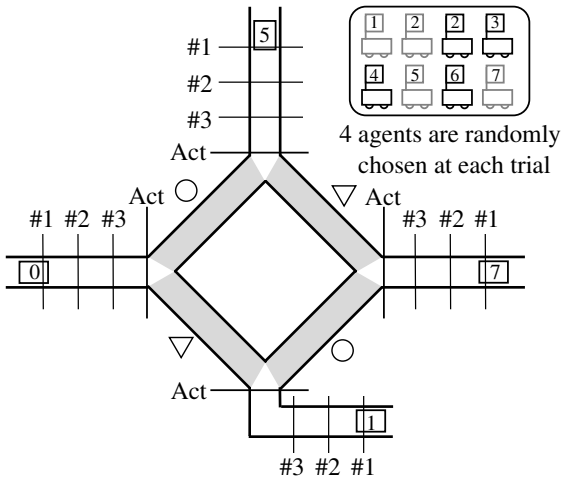


Fig. 8　Four-agent negotiation problem.

ばれた２エージェント間でコミュニケーションを通して相手との順序関係を知り，その関係だけで衝突の回避ができた．しかし，４エージェントの場合，他の一つのエージェントとの間で衝突が回避できても，それによって逆に，それ以外のエージェントとの間で衝突が起きる可能性もある．さらに，どの４つのエージェントが選ばれるかだけでなく，他の３つのエージェントが相対的にどのような位置関係で配置されるかも考慮する必要がある．したがって，われわれが各エージェントのコミュニケーション戦略を設計するのは容易ではない．

Fig. 8 に，４エージェント交渉問題の設定を示す．全部で８個のエージェントの中から４つが毎回ランダムに選ばれ，４つの入り口のどれかにランダムに置かれる．エージェントは，－１か１のコミュニケーション信号を前の問題と同様に３回送ることができる．各エージェントは，一度に４個のコミュニケーション信号を受け取る．１つは自分自身の信号，２つ目は向かいのエージェントから，３つ目はルート○側の隣から，４つ目はルート▽側の隣からの信号である．ただし，どのエージェントが選ばれているか，選ばれたエージェントがどこに配置されているか，どの信号がどこから来た信号かといった情報は知らされない．そして，最後に，○か▽のどちらのルートをとるかを決定する．そして，隣のエージェントと衝突すると罰 ($r = -5$) が，衝突しないで通り抜けると報酬 ($r = 1$) が与えられる．前述のように，得られた強化信号は，他のエージェントと配分しない．エージェントの構成は，Fig. 2 とほぼ同じであるが，コミュニケーション信号の入力の数が２から４に増える．この問題は，選ばれたすべてのエージェントが，Fig. 8 の菱形のルートを時計回りか反時計回りに回る場合が解となる．そのためには，向かい側のエージェントとは同じで，隣とは違うルートをとれば良い．

3.2.2　結果

この問題で解ける最大のエージェント数は８より大きく，解には様々なバリエーションがあった．ある解では，２エージェント問題の Table 1 でのエージェント０のように，特定のエージェントが相手や場所によらず常に同じコミュニケーション信号と行動をとった．しかし，ほとんどの場合，すべてのエージェントが状況に応じてルートを変化させていた．

この問題も，解ける確率は少ないものの，各エージェントの中間層ニューロン数を２個としても解を得ることができた．このときの解の詳細を述べる．Table 2 に，典型的な交渉の例を示す．Table 2 (a) では，エージェント０，２，３，７が選ばれている．最初と２回目のコミュニケーション信号は４つのエージェントとも－１で同じであるが，エージェント７だけが，３回目に信号を１に変化させた．そして，最終的に，エージェント０と３が行動１ (ルート) をとり，エージェント２と７が行動－１ (ルート) をとった．Table 3 に，学習後にすべての組み合わせで交渉を行った場合の各コミュニケーション信号と行動の平均値とコミュニケーション・行動のパターンの数を示した．２エージェント問題の際の Table 1 のようなエージェントのソートはここでは行っていない．これらの値の符号は，ルート○かルート▽かの好みを表し，絶対値がその好みに対する固執度を表している．最初のコミュニケーション信号は，常に一定であるため，残りの 2 回のコミュニケーション機会と行動選択の計 3 回で，とりうるすべてのパターンは $2^3 = 8$ 個であるが，エージェント０は，こ

Table 2　Typical negotiation patterns after successful learning in four-agent negotiation problem

	#1	#2	#3	Act
Agent 0	-1	-1	-1	1 ○
Agent 2	-1	-1	-1	-1 ▽
Agent 3	-1	-1	-1	1 ○
Agent 7	-1	-1	1	-1 ▽

(a)

	#1	#2	#3	Act
Agent 2	-1	-1	-1	1 ○
Agent 7	-1	1	1	-1 ▽
Agent 0	-1	-1	-1	1 ○
Agent 5	1	1	1	-1 ▽

(b)

	#1	#2	#3	Act
Agent 2	-1	-1	-1	1 ○
Agent 5	1	1	1	-1 ▽
Agent 7	-1	-1	-1	1 ○
Agent 4	1	1	1	-1 ▽

(c)

	#1	#2	#3	Act
Agent 2	-1	-1	-1	1 ○
Agent 4	1	1	1	-1 ▽
Agent 5	1	1	1	1 ○
Agent 7	-1	-1	1	-1 ▽

(d)

Table 3　The average of the communication signals and action of each agent over all combinations and the number of the communication patterns after learning. The values in the table are $2p - 1$ where p is the probability that the output is 1.

	com #1	com #2	com #3	Action	num of patterns
Agent 0	-1.0	-0.429	-0.457	0.352	7
Agent 1	1.0	0.029	0.286	-0.257	4
Agent 2	-1.0	-1.000	-1.000	0.581	2
Agent 3	-1.0	-0.086	-0.343	0.181	5
Agent 4	1.0	1.000	1.000	-0.571	2
Agent 5	1.0	0.771	0.533	-0.486	5
Agent 6	1.0	0.429	0.324	-0.286	5
Agent 7	-1.0	-0.771	-0.552	0.486	4

の８パターンのうち，７個のパターンをとっており，非常に柔軟で適応的である．また，すべてのエージェントにおいて，３回のコミュニケーション信号の好みの符号は等しく，行動の好みの符号は反対である．このことから，この解では，コミュニケーション信号－１が行動１ (ルート○) の主張であり，コミュニケーション信号１が行動－１ (ルート▽) の主張であると解釈できる．エージェント間のすべての組み合わせでの交渉結果を見ると，つぎのようないくつかのルールを見つけることができた．

（ケース１） 選ばれた４つのエージェントの好みのルートがすべて○の場合，つまり，エージェント０，２，３，７が選ばれた場合（Table 2 (a)），エージェントの配置によらず，コミュニケーションの機会＃３でエージェント７がコミュニケーション信号を変化させ，最終的に向かいのエージェントと共にルートをとった．

（ケース２） Table 2 (b) のように，３つのエージェントの好みのルートが○で，他のエージェントの好みのルートが▽である場合，▽を好むエージェントおよび隣接する２つのエージェントは意志を変えず，向かいにいるエージェントが意志を変えてルート▽を選択した．

（ケース３） ４つのエージェントの好みが半々の場合，さらに２つの場合に分けられる．１つは，同じ好みのエージェントが向かい合っている場合で，この場合は，全エージェントが最初の意志を貫いた (Table 2 (c))．

（ケース４） 好みが半々に分かれた場合で，同じ好みのエージェントが隣同士の場合は，エージェントのコミュニケーションと行動は複雑で，簡単なルールを見つけることができなかった (Table 2 (d))．

ケース１とケース２のそれぞれと対称，つまり，４つのエージェントの好みがすべて▽の場合や，3 つのエージェントが▽を好み，一つのエージェントが○を好む場合もあるが，ちょうどケース 1 および２とそれぞれ対称的な戦略であった．ニューラルネットの初期値を変えた他のシミュレーションでも，同様に８個のエージェントの好みは半々になるという対称性を予測していた．しかし実際には，最も非対称な場合で，２つのエージェントがルート○を好み，残りの６個がルート▽を好むという結果になった場合もあった．

Fig. 9 に前述の特徴的なエージェント０，２，７のニューラルネットの結合の重み値を示す．各エージェントの特徴はつぎのように発現することがわかる．

[全エージェント] すべてのエージェントで，入力１および４から中間層への結合の符号はほぼ同じで，入力２および３からの結合の符号はその逆になっている．これによって，向かいとは同じ行動を，隣とは違う行動を得る力が働く．

[エージェント０] 全般的に，他のエージェントのコミュニケーション信号の入力である入力１，２，３から中間層への結合の重み値の絶対値が大きく，逆に，文脈入力（中間層からのフィードバック入力）および自分のコミュニケーション信号入力から中間層への結合の重み値の絶対値は小さい．つまり，エージェント０は自分の意志よりも，他のエージェントの意向を優先してコミュニケーション，行動を決定している．したがって，相手によって様々なパターンが生成されると考えられる．

[エージェント２] 中間層からコミュニケーション出力への結合の重み値が０に近いため，エージェント０とは逆に，他のエージェントの意向は参考にせず，コミュニケーション信号を生成していることがわかる．このことは，Table 3 からもわかる．

[エージェント７] Table 2 の (a)(d) のように，最初の全エージェントのコミュニケーション信号が２回目と全く同じでも，２回目と３回目に違う信号を生成することができる．これも，Fig. 9 (c) の結合から大まかに説明できる．

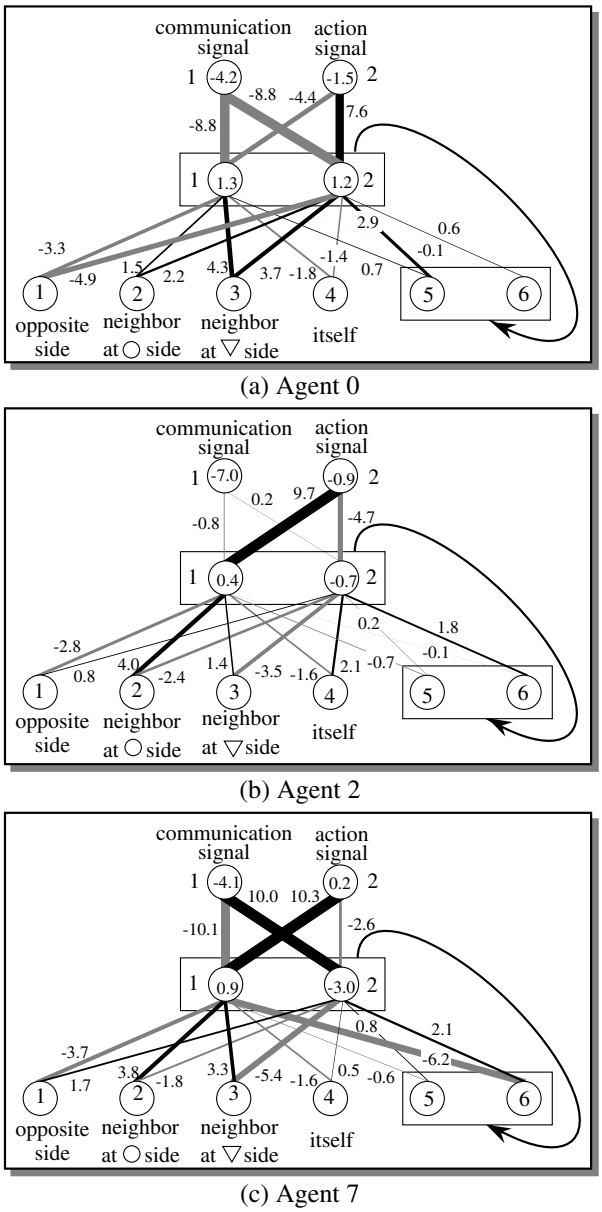


Fig. 9 Connection weights of the neural network after learning of the four-agent negotiation problem.

中間層ニューロン２はバイアス値が小さいため，＃１では出力値が非常に小さい．したがって，＃２では，入力６からの負の結合により中間層ニューロン１の値が大きくなり，コミュニケーション出力は小さくなる．しかしこのとき，入力３からの負の結合によって中間層ニューロン２の値は大きくなり，＃３では，中間層ニューロン１の値が小さくなってコミュニケーション出力が１に近くなる．

中間層を４個にして，ニューラルネットの結合の初期値を変えて 100 回シミュレーションを行った．BPTT 適用時には 42 回で解が求まり，通常の BP では，26 回で解が求まった．このことから，BPTT のほうが多少優位であったものの，通常の BP でも解が求まることがわかった．

4. 結　　論

マルチエージェント間でのコミュニケーションを，観察した情報の伝達と自分の意志の伝達の２つに大きく分類した．そして，後者の場合，交渉のための双方向のコミュニケーションがループをなすことによって，ダイナミクスが生まれることを示し，これをダイナミックコミュニケーションと呼んだ．そして，各エージェントがリカレントニューラルネットを用いて強化信号から自律的にダイナミックコミュニケーションを学習するアーキテクチャを提案した．

例題の衝突回避交渉の問題を通して，エージェントにコミュニケーションの内容やその表現法および戦略を予め与えなくても，行動後の報酬や罰だけを用いた簡単な学習によって，コミュニケーション戦略に個性が発現し，複雑かつ適切なコミュニケーションを通して，衝突を回避できることがわかった．このとき，リカレントニューラルネットが必要に応じて過去の情報を保持しながら，相手のエージェントに対して適応的に交渉を進めていることがわかった．また，各エージェントのニューラルネットが，単にテーブルルックアップのように入力と出力のマッピングを学習しただけでなく，解という固定点に収束するダイナミクスを学習したと考えられる．これらの過程は，自律分散的な変化によって系のポテンシャルが落ちていくという意味で，ホップフィールドネット [10] を使った最適化の過程に類似している．また，過去の履歴を入力として与える方法（バッファ方式）と比較して，リカレントネットを用いたほうがよりその傾向が強いことがわかった．また，１対１の交渉だけではすまない４エージェント間の交渉の問題においても，強化信号を他のエージェントと分け合うことなく，全エージェントが満足する解を得ることができた．さらに，リカレントニューラルネットの学習アルゴリズムとして，通常の BP と BPTT の２つを比較したが，あまり大きな差は生じなかった．

謝辞

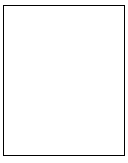
本研究の一部は，日本学術振興会未来開拓学術研究推進プロジェクト”生物型適応システム” (JSPS-RFTF96I00105) および文部省科学研究費基盤研究 (B) (No. 10450165) の補助のもとで行われた．ここに謝意を表する．また，有用なコメントを頂きました校閲者の方に，感謝致します．

参 考 文 献

1）Werner, G. M. & Dyer, M. G., ”Evolution of Communication in Artificial Organisms”, Proc. of Artificial Life II, pp.1-47 (1991)
2）Nakano, K., Sakaguchi, Y., Isotani, R. & Ohmori, T., ”Self-Organizing System Obtaining Communication Ability”, Biological Cybernetics, 58, pp.417-425 (1988)
3）Davis, R. & Smith, R. G., ”Negotiation as a Metaphor for Distributed Problem Solving”, Artificial Intelligence, Vol. 20, No. 1, pp. 63-109 (1983)
4）Kreifelts, T. & von Martial, F., ”A Negotiation Framework for Autonomous Agents, Demazeau, Y. & Muller, J.-P. eds., Decentralized A. I. 2, pp.71-88 (1991)
5）上田雅英，柴田克成，伊藤宏司，“マルチエージェント系における個性・社会性の学習的生成”, 計測自動制御学会第 11 回自律分散システムシンポジウム (1999)
6）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”, 計測自動制御学会第 25 回知能システムシンポジウム (1998)
7）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”, 計測自動制御学会第 11 回自律分散システムシンポジウム (1999)
8）Elman J. L., ”Finding Structure in Time”, Technical Report CRL 8801, Center for Research in Language, Univ. of California, San Diego (1988)
9）Rumelhart, D. E., Hinton, G. E. & Williams, R. J., ”Learning internal representations by error propagating”, Parallel Distributed Processing, Vol. 1, MIT Press, pp. 318-362 (1986)
10）Hopfield, J. J., ”Neural Networks and Physical Systems with Emergent Collective Computational Abilities”, Proc. of Natl. Acad. Sci. USA, Vol. 79, pp.2554-2559 (1982)

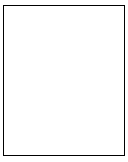
［著　者　紹　介］

柴　田　克　成



1989 年 東京大学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 東京大学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 東京大学先端科学技術研究センター助手．1997 年 東京工業大学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）．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））．

伊　藤　宏　司（正会員）



1969 年 名古屋大学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 東京工業大学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）．