コミュニケーション相手の取捨選択による個体知能の効率的発達

木島康隆

2010年1月29日

目 次

第1章　序論

1.1　はじめに

旧来，ロボットの適用範囲としては，工場のラインが主であり，単純な作業を行なうことが主な用途であった．時代が進むとともに，ハードウェア技術が発達しロボットが高機能になってくると，ロボットの使用環境・用途も変化してきた．近年では，工場のみならず，家庭環境・災害現場・宇宙や深海などの極限環境といったようにロボットの適用範囲は大きく広がってきている．また，使用用途も多種多様なものになってきた．それに応じて上記のような環境で利用することを目的としたロボットの研究・開発も活発になっている [1]-[?]．こうしたロボットの動きの設計という視点でみると，旧来の工場のような環境ではある程度周囲の環境がロボットのために整備されており，環境が変化する要因は少なく，変化があまり起きない．このような環境は静的環境と呼ばれる．静的環境では，ロボットが直面する状況を設計者が想定し，それに応じた行動を設計することが可能である．対して，家庭環境，災害現場，極限環境といった環境では，環境が変化する要因は無数に存在し，常に変化している．例えば，家庭環境では，家具の位置やテレビリモコンや食器のような小物の位置といったものは，常に同じ場所にあるわけではなく，家庭に住む人間によって変わっていく．このような環境は動的環境と呼ばれ，ロボットが直面する状況すべてをを設計者が想定するのは不可能である．近年，ロボットがこうした動的環境下での動作が望まれる中，環境に適した動きの設計は大きな問題となっている．

そこで，ロボットに自律的に環境に適した動きを獲得し，行動することが要求され，盛んに研究が行なわれている．こうした研究の中の１つのアプローチとして，機械学習をロボットに組み込むことでロボットに環境に適した行動を自律的に学習させるというものがある．

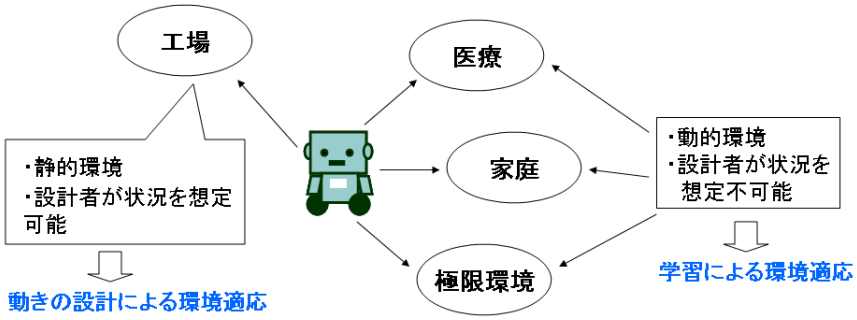

図 1.1: ロボットの接する環境の多様化と問題

1.2　機械学習

人間は過去に行なった問題とよく似た問題を解決するとき，以前に解いた経験を基によりうまく解決できる学習能力を持っている．このような学習能力をロボットに持たせることを目的とする研究は，機械学習（machine leaning）と呼ばれる．一般に機械学習は,「教師あり学習」,「教師なし学習」,「強化学習」の 3 つに分けられる．

1.2.1　教師あり学習

人間がゴルフやテニスのスイングフォームを学習するとき，上手い人のスイングを見て自身のスイングを修正していく．これによって自身の動きを上手い人のそれに近づけていく．このとき上手い人の動きは教師として機能している．教師あり学習は，理想的と考えられる出力信号（教師信号）が学習者に与えられ，学習者は教師信号を基に自身の出力信号を教師信号に近づけるように学習を行なう．教師あり学習の主なものとしてはニューラルネットワーク [17] がある．

教師あり学習は，学習のための教師信号が適切でなければ，学習結果も適切でないものになってしまう．これは，先の例で教師を下手な人間にしてしまえば，学習者の動きも下手なものになってしまう，ということである．ロボットの場合，学習者に教師信号を与えるのは人間である．人間がロボットが直面する環境を予測し，それに応じて教師信号を設計する．これは，予測が容易な静的な環境であればよいが，動的環境下ではロボットが直面する環境の予測が非常に困難であるため，人間による教師信号の設計ミスや，そもそも設計できないといったことが発生することが考えられる．こういった場合，教師あり学習を適用するのは難しい．

1.2.2　教師なし学習

地上には様々な動物がいるが，脚が４本のものや６本のもの，草を食べるものや動物を食べるものがいて，色や形，大きさなどの特徴から，どうやら動物はいくつかの種類があるんだ，ということは人間だれでも理解できる．これを動物図鑑を見ながら，これはイヌ，これはサル，と学習していくのが教師あり学習であり，そういう手本はなくても，多数のサンプルの相関や統計的な偏りをもとに，それらをグループ分けしたり，特徴量のベクトルに分解したりするのが教師なし学習であり，自己組織化と呼ばれることもある．教師なし学習には，主成分分析・独立成分分析 [18] などのように信号を特徴的な成分に分解するもの，混合正規分布モデル [18]，自己組織化マップ [19] などのように離散化しクラス分けするものの２つがある．

1.2.3　強化学習（reinforcement learning）

強化学習 [] は，報酬をもとにした学習である．学習者は得られる報酬を最大化するように学習を行なう．たとえば，犬におすわりを教えるときを考える．犬が飼い主の指示に従っておすわりをしたらご褒美としてエサを与えるようにして，訓練を行なう．すると，次第に犬は試行錯誤を通して飼い主の指示に従っておすわりをするように学習する．このとき，犬に与えるエサが報酬となり，犬はエサ（報酬）を得られるように試行錯誤しながらおすわりの状態に至る行動（後ろ足をたたむ）を学習する．このように，目標となる状態（例ではおすわりしている状態）に対して報酬を設定しておけば，その状態に至るまでの行動を試行錯誤で学習を行なうのが強化学習の特徴である．また，報酬が教師信号になる教師あり学習ともとれるが，報酬に至るまでのプロセスは教師信号には含まれていないため，厳密には教師あり学習ではない．ここでは，強化学習の基礎となる概念について説明したが，詳しい説明は第２章で行なう．

1.3　機械学習をロボットに適用する際の問題点

ロボットの行動学習は，認識される周囲の状況に対して適切な行動を探索することである．周囲の状況は，ロボットに搭載されているセンサを介して認識される．また，行動は，アク

チュエータを介して行なわれる．ロボットの学習は，周囲の状況を認識する能力と行動能力が作る空間内から状況に適した行動を探索することである．したがって，センサーの高度化または搭載数の増加により，ロボットの状況認識能力が向上し，認識できる状況が増加すると，それだけそれぞれの状況に適した行動を探索するのに時間が掛かってしまう．同様に，ロボットのアクチュエータの増加によって，取れる行動数が増加すると，それぞれの状況にあった行動の探索に時間が掛かる．近年，ハードウェア技術の発展によってヒューマノイドロボットに代表されるような，多センサ多アクチュエータのロボットの開発が盛んになっている．このようなロボットは高環境認識能力，高行動能力を持っており，これらのロボットが実環境で学習を行なうことを考えた場合，その高い能力ゆえに膨大な学習時間が掛かってしまうことが予想される．実環境では，学習にいつまでも時間を割くことはできず，すばやい環境適応が求められる．よって，このようなロボットの高度化による学習時間の増加は問題である．

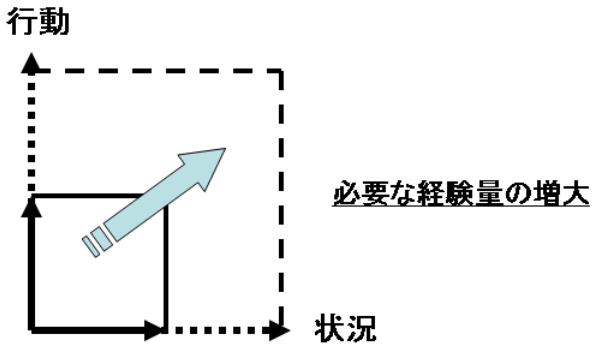

図 1.2: センサとアクチュエータの種類の増加による学習空間の増大

1.4 群を用いた個体知能の発達に関する研究

人間を始めとする高度な生物は，個々で学習を行い知能を発達させるだけではなく，他の個体とコミュニケーションを行い情報をやり取りすることで自身が未だ知らない知識を獲得し，それを利用し自身に役立てることができる．例えば，今まで行ったことの無い場所に行く時を考えると，自分ひとりでは標識や地図を頼りにたどり着くことが可能である．しかし，土地勘のある人に道を尋ねた方がより早くたどり着くだろう．このように他者とコミュニケーションを行うことで，個体単体が入手できる情報量を増やしより効率的に知能を発達させる研究が行なわれている [20]〜[21]．私は文献 [23]，[24] でコミュニケーションを利用した個体学習促進システムを提案した．このシステムでは，コミュニケーション情報として，現在使っている学習手法とその学習手法を用いて得られた報酬を採用した．エージェントは，コミュニケーションを通じて自身が直面する環境に適した学習手法を学習するとともに，その環境下で経験する個々の状況に適した行動を学習する．実験タスクとして，N 本腕バンディット問題に適用し，提案手法の有効性を示した．

1.5 群を用いた個体知能の発達に関する研究の問題点

群を用いた個体知能の研究では，どの個体とコミュニケーションを行うかといったコミュニケーション個体の設定は，ロボットの設計者によって定められてきた．このとき，設計者はロボットに対して設計者がロボットの特性（目的・身体構造等）に合ったコミュニケーション相手を設定する．このため，ロボットにとって悪影響を与えるコミュニケーション相

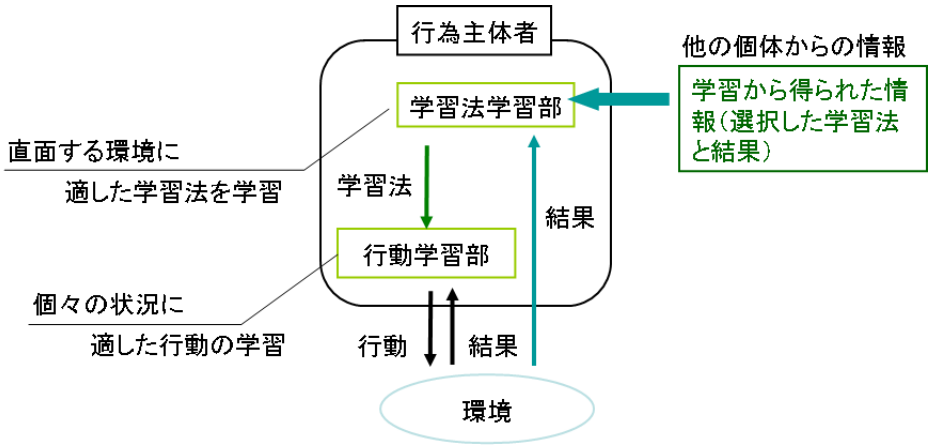

図 1.3: 学習手法をコミュニケーション情報とした個体知能発達促進システム

手を設定される可能性があり，結果として知能の発達に悪影響を与えることが考えられる．また，群の中に新たな個体が追加された場合，群の中のロボットが故障により使用不可能になった場合には新たにコミュニケーション個体の設定をやり直す必要が出てくる．設計者が把握できる程度の規模の群であれば，設計者の手によって設定し直すことも可能であるが，設計者が把握出来無い規模の群を構築している場合，新たにコミュニケーション個体を設定するのは難しい．このような問題は，災害現場での被災者の探索・救助のような多数台のロボットが利用され，かつ故障が起こりやすい環境では重要となってくる．そこで，ロボットが自律的にコミュニケーション相手を取捨選択することが望ましい．

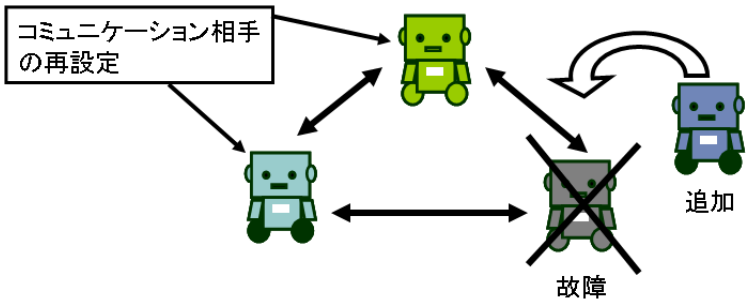

図 1.4: 従来研究の問題点

1.6 本研究の目的

本研究では，ロボットが自律的にコミュニケーション相手を選択するシステムを提案する．多くの自身に有益な情報を持つものとコミュニケーションしより効率的に知能を発達させることできる．

1.7 アプローチ

コミュニケーション相手の取捨選択は，コミュニケーションを通して経験的に行なう．他者から入手される情報が有益なものならば，それに基づいて行動した結果は良くなり，逆に不利益なものであれば結果は悪くなると考えられる．そこで，自身にとって良い結果になるような情報を提供する他者は高く評価し，そうでないものについては低い評価を行なうこと

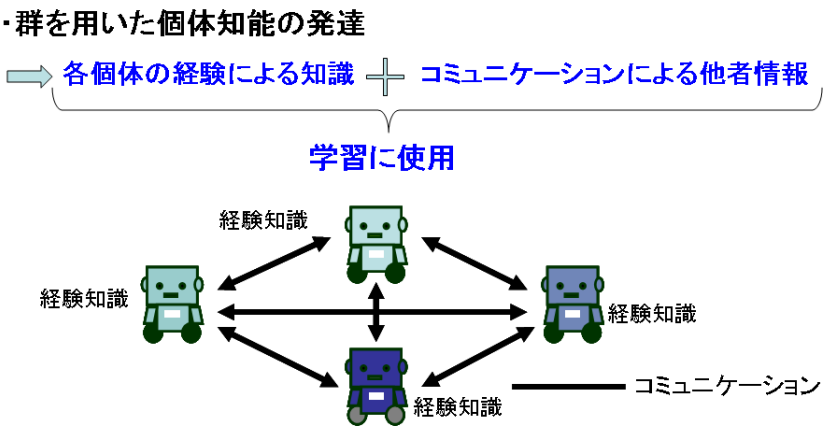

図 1.5: 本研究の目的

でコミュニケーション個体の取捨選択を行なう．評価の高い他者とコミュニケーションすることで，自身に有益な情報を取り入れ効率的に個の発達を実現する．

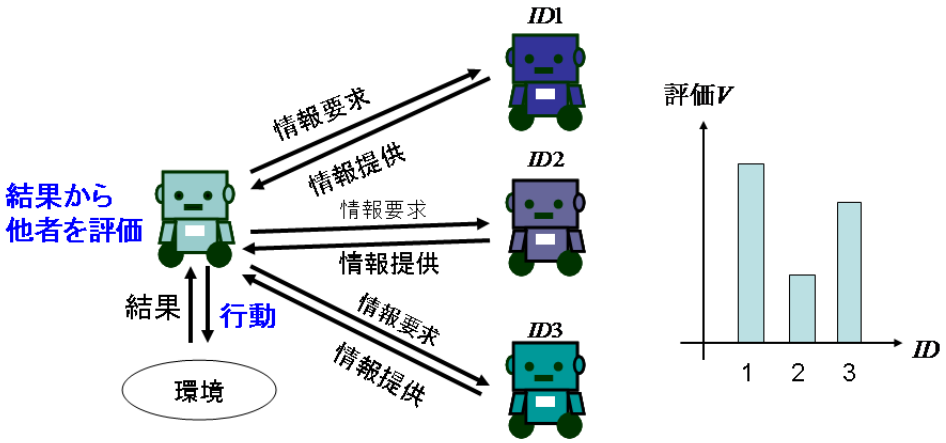

図 1.6: アプローチ

1.8 本論分の構成

ここでは本稿の構成を示す．

第 1 章 本論文の背景と目的，アプローチについて述べる．

第 2 章 本研究で，扱う強化学習について述べる．

第 3 章 報酬遅延の起こらない環境に対する手法を解説し，実験を行なう．

第 4 章 報酬遅延の起こる環境に対する手法を解説し，実験を行なう．

第 5 章 本研究のまとめと今後の課題について述べる．

謝辞・参考文献 謝辞・参考文献について述べる．

第2章　強化学習

本章では，本論文で用いる機械学習手法である強化学習 [25][26] について概要と今回実験で用いた学習について述べる．

2.1　強化学習の概要

2.1.1　環境とエージェントとの相互作用とエージェントの目的

強化学習では，エージェント（学習者）は周囲の環境状態（以降，状態）を認識し，その状態で目的を達成するためには何をすべきかを学習する．学習は，行った行動の良し悪しを数値化したものである報酬を基にして行われる．報酬は環境によって与えられる．エージェントはどの行動がより高い報酬に結びつくかを探索し，得られる報酬の総和を最大化することを目的とする．学習の流れを以下の箇条書きに示す．

1. エージェントは時刻 t でセンサを通して知覚される環境の状態 s_t に基づいて意思決定を行い，行動 a_t をとる
2. エージェントの行動 a_t の結果として環境から報酬 r_t を受取る
3. エージェントの行動 a_t により，環境は状態 s_{t+1} へ遷移する

エージェントは環境に対してこのような状態の観測，行動，状態の変化，報酬獲得の獲得という一連の流れを繰り返す（環境との相互作用）ことで学習を行う（図 2.1）．

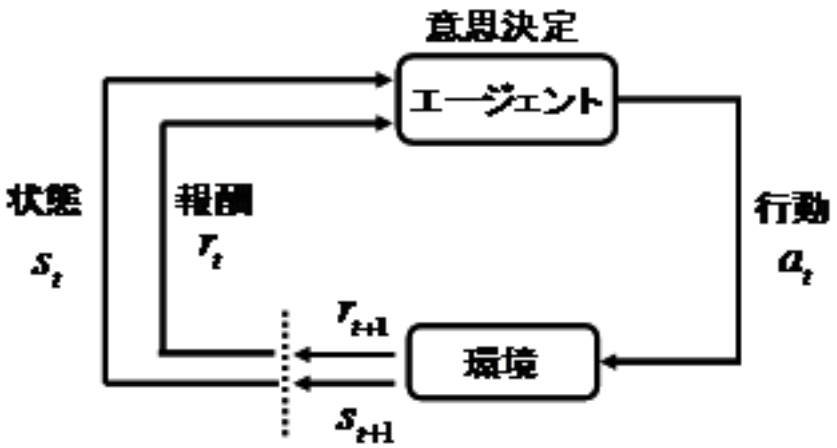

図 2.1: 環境との相互作用

先に述べたようにエージェントの目的は得られる報酬の総和を最大化することである．報酬はエージェントの設計者が設定する．したがって，設計者の目的を達成するような行動に対して高い報酬を設定しておくことで，目的に至るまでの行動のプロセスはエージェントが自動的に獲得する．

2.1.2 強化学習の特徴

強化学習の特徴としては以下のようなことが挙げられる．

- 報酬を基にした探索
- 遅延報酬に対応

報酬を基にした探索　強化学習では，学習のための正解情報を直接与えられることはない．その代わりに，エージェントは行った行動に対してその行動の良し悪しを報酬として与えられる．与えられた報酬が最良か最悪かはエージェントは知らされない．したがって，エージェントはどの行動がより高い報酬に結びつくかを試行錯誤を通して探索する．

遅延報酬に対応　強化学習では，試行錯誤を通して最終的に目的を達成した際に報酬が与えられることが多く，ある時点でエージェントが選択した行動の良し悪しの判定には時間的な遅れが存在する．このことを遅延報酬と呼ぶ．

2.1.3 応用上されていること

文献 [27] によると以下のようなことが応用上期待できる．

制御プログラミングの自動化・省力化

環境に不確実性や計測不能な未知のパラメータが存在すると，タスクの達成方法やゴールへの到達方法は設計者にとって完全にはわからない．よってロボットに対してタスクを遂行するための制御規則をプログラムすることは，設計者にとって非常に負担になる．一方，達成すべき目標を報酬によってロボットに明示することは前記に比べれば遥かに簡単である．そのため，タスク遂行のためのプログラミングを強化学習で自動化することにより，設計者の負担軽減が期待できる．また，十分に優れた性能を持つ強化学習エージェントをコントローラとして１つだけ開発しておけば，あとはロボットの目的に応じて報酬の与え方だけを設計者が設定するだけで，あらゆる種類のロボット制御方法を同一のコントローラによって自動的に獲得できる．

ハンドコーディングよりも優れた解

試行錯誤を通じて学習するため，人間のエキスパート（専門家）が得た解よりも優れた解を発見する可能性がある．特に不確実性（摩擦やガタ，振動，誤差など）や計測が困難な未知パラメータが多い場合，人間の常識では対処し切れないことが予想され，強化学習の効果が期待できる．この新しい解の発見には２つのアプローチが存在する．１つは，エキスパートの制御規則を学習初期状態に設定して，それを改善する方法である．そしてもう一つは，全くのゼロから学習を開始し，設計者にとっては意外な新しい解を発見する方法である．

自律性と想定外の環境変化への対応

機械故障などの急激な変化やプラントの経年変化のような緩慢な変化など，予め事態を想定してプログラミングしておくことが困難な環境の変化に対しても自動的に追従する．特に宇宙や海底など，通信が物理的に困難な極限環境の場合や，通信ネットワークの制御のように現象のダイナミクスが人間にとって速すぎる場合において，強化学習の自律的な適応能力が特に威力を発揮する．

2.1.4　強化学習の構成要素

ここでは強化学習の構成要素である，環境・行動学習手法（方策）・報酬関数・評価関数・行動学習手法について解説する．

・環境

環境はエージェントがセンサを通して知覚することができる全ての状態を内包している．例として，家庭環境を考える．家庭環境には机や椅子のような家具，リモコンや食器といった小物類といったものが存在する．これらの家具や小物は置いてある場所が変わったり，経年劣化によって外見がくすんでくるなど状態が変化している．このように家庭環境は状態の変化を全て保持していると捉えることができる．エージェントはセンサを介して自身が直面している環境の一部を認識する（2.2）．エージェントのセンサには認識できる環境の範囲が存在する．エージェントはセンサの認識できる範囲の環境の状態しか認識することができない．したがって，エージェントが認識する状態は環境の一部を切り取ったものになる．このセンサを介して認識できる一部の範囲のことを状態と呼ぶ．また，環境はエージェントの行動の結果，どのように状態が遷移するかといったルール情報も保持している．したがって，エージェントある状態である行動をとった時に次に遷移する状態が予測できる．

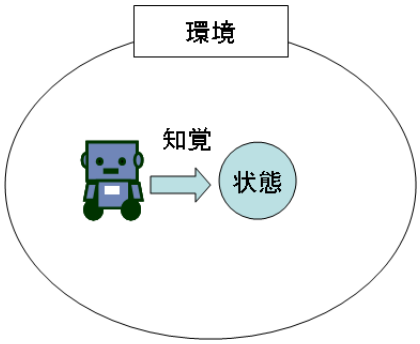

図 2.2: 環境と状態

・行動選択手法（方策）

行動を決定する手法である．例を挙げる．ある状態で行動を決定する際にその行動をランダムに決定する手法，今までで最も高い報酬を得られた行動に決定する手法，過去の報酬の累積から確率的に決定する手法，というように行動の決定のルールが行動選択手法である．

・報酬関数

報酬関数は強化学習問題において目標を定義する．目標は設計者がエージェントに学習させる状態や行動である．この関数は状態行動対を報酬という数値情報として出力する．報酬は現在の状態に備わった望ましさを表している．ゆえに，報酬関数は即時的な意味合いでエージェントにとってなにが良いのか示している．一般に報酬関数は設計者が設計するものでエージェントが変更することはしない．

・価値関数

報酬関数が即時的な意味合いで何が良いのか示しているのに対して，価値関数は，最終的な状態または行動の価値を決定する．価値とは，エージェントがその状態を基点として将来にわたって入手できる報酬の期待値である．報酬はその環境が即時的で固有の望ましさを決定するのに対して，価値はその後に続きそうな状態群とそれらの状態群で得られそうな報酬を考慮に入れた上での長期的な望ましさを示すものである．例えば，ある状態では常に低い報酬しか得られないかもしれないが，高い報酬が得られるような状態が規則的にそれに続くのならば，高い価値を持つ．

・行動学習手法

価値関数は報酬関数を基に更新されてゆく．行動学習手法は報酬関数をもとに価値関数を更新する手法である．例えば，過去に得られた全ての報酬の平均するといったように，どのように価値関数を更新するかを規定したものが行動学習手法である．

2.1.5 強化学習の流れ

強化学習の流れを図 2.3 に示す．環境から知覚した状態 s によって，エージェントは自身が行うことのできる行動の中から，その状態 s における行動価値 Q に基づき行動選択手法を用いて行動 a を選択（意思決定）し，実行する．その結果，環境より得られた報酬 r を基に，エージェントは状態 s において選択した行動 a の価値 $Q(s,a)$ の更新を行動評価手法によって行い（学習），次回同様の状態においての行動選択に生かす．エージェントは，行動価値 Q を基にして行動選択手法を用い行動を決定する．行動の結果受け取る報酬を基に行動価値 Q を行動学習手法により更新する．行動の決め方を規定する行動選択手法と行動価値の更新の仕方を規定する行動学習手法によって学習の性質を決定する．したがって，強化学習の学習法は行動選択手法と行動学習手法の 2 つの組み合わせによって決定する．

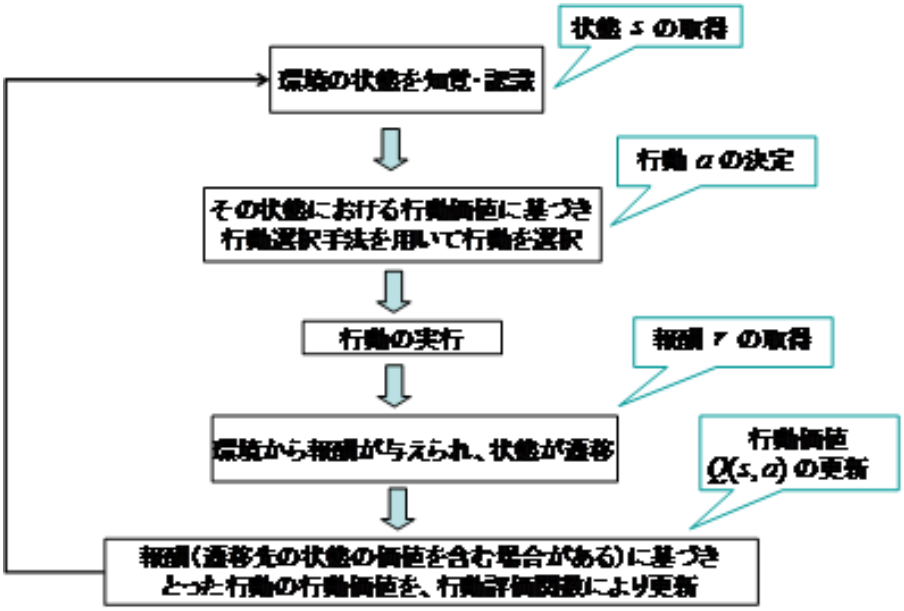

図 2.3: 強化学習の流れ

2.2 行動選択手法

強化学習における行動選択手法とは，エージェントが認識した状態 s においてとる行動 a を選択する際に用いられる手法である．ここでは，本実験で用いる主な行動選択手法として，greedy 法， ε - greedy 法，softmax 法について述べる．

2.2.1 greedy 法

最も高いと推定された行動価値を持つ行動（あるいは行動群から 1 つ）を選択する（図 2.4）．この方法は常に即時の報酬を最大にするために，現在の行動価値を利用するものである．すなわち，価値が低いと判断される行動に対しては，それが本当はさらに良いかもしれないという可能性を確かめる目的での試行を一切行わないという性質がある．

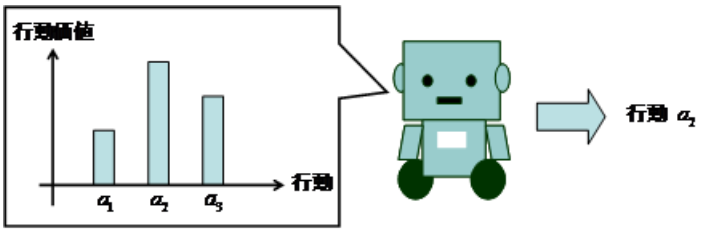

図 2.4: greedy 法

2.2.2 ϵ-greedy 法

ϵ-greedy 法は，基本的には推定される行動価値が最も高い行動（グリーディな行動）を選択するが，たまに小さい確率 ϵ で行動価値の高さとは無関係にランダムで行動を選択する手法である（図 2.5）．常に行動評価値の最も高い行動しか行わない greedy 法とは異なり，確率 ϵ で探査行動を行う．しかし，確率 ϵ における行動選択の際にほとんど最悪と思われる行動とほとんど最適に近い行動を選択する可能性が同じくらいの高さになるという欠点がある．

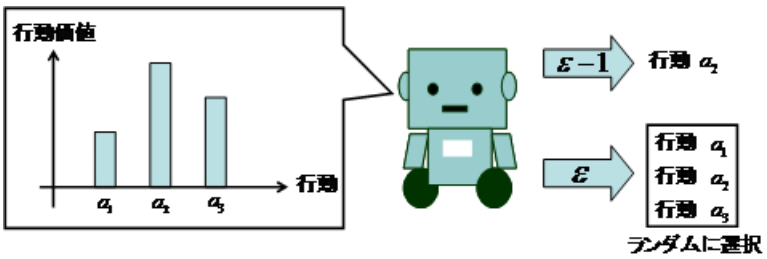

図 2.5: ϵ-greedy 法

2.2.3 softmax 法

softmax 法は，行動価値を等級付けした関数によって行動確率を変化するやり方である．すなわち，行動価値の最も高い行動には最も高い選択確率が与えられ，他のすべての行動は，その推定価値に従って重みをかけられ，ランク付けされる（図 2.6）．Softmax 法では一般に，Gibbs 分布 [33]，あるいは Boltzmann 分布 [34] が使われる．具体的には，t 回目のプレイにおける行動 を選択する確率 は式 (2.1) で与えられる．ここで，$\pi_t(s, a)$ は時間 t，状態 s

で行動 a を選択する確率，$Q_t(s,a)$ は時間 t，状態 s で行動 a を選択したときの行動価値である．τ は温度と呼ばれるパラメータでこの値の大小で重みのつけ方が変わる．τ を小さくすると各行動推定価値の差が少しでも行動選択確率は大きく異なる．逆に，τ が大きいと各行動の推定価値の差が大きくても行動確率の差は小さくなる．つまり，τ が小さいと確定的になり，τ が大きいと確率的になる．τ の大小の基準は報酬の大きさとに依存するため，τ はタスクごとに設計者が綿密に定めることが望ましい．

$$\pi_t(s,a) = \frac{e^{Q_t(s,a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(s,b)/\tau}} \tag{2.1}$$

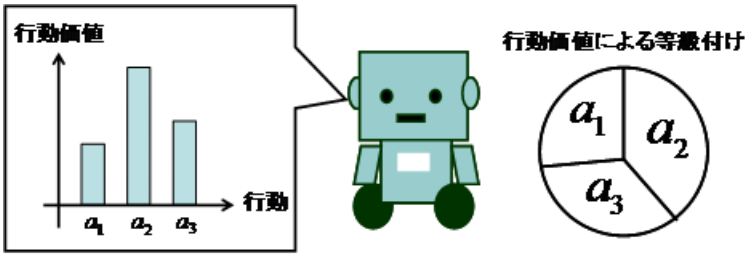

図 2.6: softmax 法

2.3 行動評価手法

強化学習において，エージェントは行動の真の価値そのものを知ることはできないため，毎回の行動によって得られる報酬からその行動の真の価値を推定する．そして，その推定値を使って行動選択手法を通して行動を選択する．この行動の真の価値を推定するための方法が行動評価手法である．ここでは，行動評価手法として，標本平均手法，加重平均手法，Q 学習法の 3 手法について述べる．

2.3.1 標本平均手法

標本平均化手法では，その行動が選ばれたとき実際に得られた報酬を平均化してゆく．時間 t，状態 s，行動 a について標本平均手法を用いた時の行動の価値 $Q_t(s,a)$ は式 (2.2) で更新する．分子は過去に状態 s で行動 a をとったときに得た報酬の総和である．また分母は状態 s での行動 a の累積選択回数である．$k_{sa} = 0$ の場合には，$Q_t(s,a)$ を $Q_0(s,a) = 0$ のような初期値にする．大数の法則より，$k_a \leftarrow \infty$ の極限において $Q_t(s,a)$ は真の価値 $Q^*(s,a)$ に収束する．

標本平均化手法は定常環境での動作に適したものである．しかし，非定常環境ではあまり有効とはいえない．これは，標本平均化手法では，より多くの試行を行なうほど結果が反映されにくくなるためである．試行回数が増えることでその行動の選択回数が増え，分母の値が大きくなる．その結果，最新の報酬が評価値に影響を与えにくくなる．このため，非定常環境下での評価関数はより最新の報酬に対し重みをおいて評価するといった工夫が必要になってくる．そのような手法が次節で紹介する加重平均手法である．

$$Q_t(s,a) \leftarrow \frac{r_{s1} + r_{s2} + \cdots + r_{sk}}{k_{s,a}} \tag{2.2}$$

2.3.2　加重平均手法

加重平均手法は，遠い過去の報酬よりも最近に受け取った報酬の方により重みを与えるような方法である．重みを与えるために，定数値のステップサイズ・パラメータを使用する．時間 t で状態 s において行動 a をとり報酬 r_t を受け取ったときの行動価値 $Q_t(s,a)$ の更新式は式 (2.3) のようになる．ここで，α はステップサイズパラメータ $(0 \leq \alpha \leq 1)$ である

$$Q_t(s,a) \leftarrow Q_t(s,a) + \alpha[r_t - Q_{t-1}(s,a)] \tag{2.3}$$

2.3.3　Q 学習法

標本平均手法，加重平均手法ともに行動の都度入手される報酬を基に評価を行なう．これは行動ごとに報酬が得られる環境に対してしか適用できないということである．したがって，迷路問題のような報酬が目標状態に対してのみ割り振られるタスク（遅延報酬）の場合，途中の経路に対し状態価値が割り振られることはない．そのため，遅延報酬タスクに対して標本平均手法，加重平均手法では学習が進まない．

Q 学習法は，遅延報酬環境下でも利用できる手法である．Q 学習では，現在の状態で選択した行動の価値とその行動の結果，推移した先の状態の行動価値によって現在の行動価値を更新する (図 2.7)．迷路問題の場合，ゴール時にもらえる報酬価値をスタートまでのルートに対し伝播させることが可能になる．これによりスタートからゴールまでのそれぞれの状態に対して，その状態の価値が算出されるため学習が可能となる．

時間 t で状態 s において行動 a をとり報酬 r_t を受け取ったときの行動価値 $Q_t(s,a)$ の更新式は式 (2.4) のようになる．ここで α はステップサイズパラメータ $(0 \leq \alpha \leq 1)$，γ は割引率 $(0 \leq \alpha \leq 1)$ である．

$$Q_t(s,a) \leftarrow Q_t(s,a) + \alpha[r_t - \gamma Q_{t+1}(s,a) - Q_t(s,a)] \tag{2.4}$$

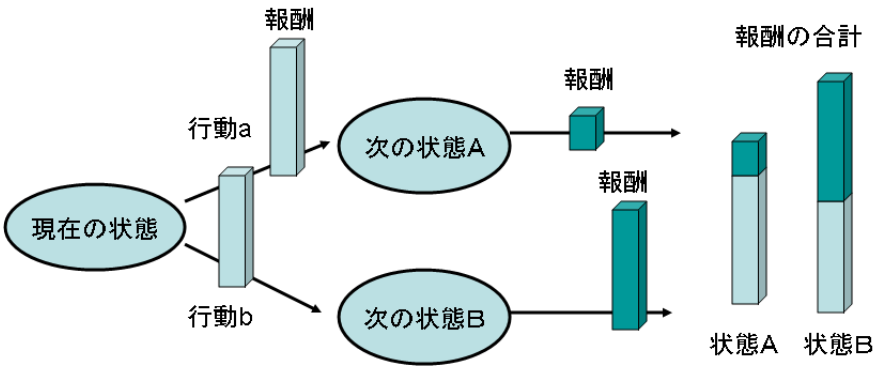

図 2.7: Q 学習法

2.4　まとめ

本章では，本論文で扱う学習法である強化学習について概要を説明した．また本実験に使用する行動選択手法，行動評価手法について基本的なものを説明した．

第3章　即時報酬環境下でのコミュニケーション相手の選択による効率的知能発達

本章では，2 章で紹介した強化学習の枠組みを用いてシステムを構築し，実験により有効性を検証する．強化学習は実ロボットに使われることが多い手法である [28]-[32]．本システムもいづれは実ロボットへの適用を考えているので本手法では強化学習を用いてシステムを構築する．本章で構築するシステムは，行動した結果すぐに報酬が入手できるような即時報酬型（行動に対する報酬が即時与えられる）の環境に対しての手法である．

3.1　作成するシステムの概要

作成するシステムの概念図を図 3.1 に示す．本システムは，コミュニケーション個体の選択を学習する部分と直面する状況に対して適切な行動を学習する部分の 2 つの学習を行う．エージェントが保持する知識は，他者に関する知識（自身から他者に対しての評価 (V)）と行動に関する知識 (Q) となる．他者に関する知識 V はコミュニケーション相手の取捨選択の際の指標なり，行動に関する知識 (Q) は直面する状況に対して行動を選択する際の指標となる．それぞれの知識は，行動の結果得られる報酬を基にして更新，蓄積される．他者からの情報は，自身の知識とともに行動選択に利用することで自身の保持する知識量を増やし，行動選択に反映する．エージェントは自身にとって適切なコミュニケーション相手を学習すると同時に直面する状況に適した行動を学習する． 強化学習の適用部は図 3.2 の赤い枠で

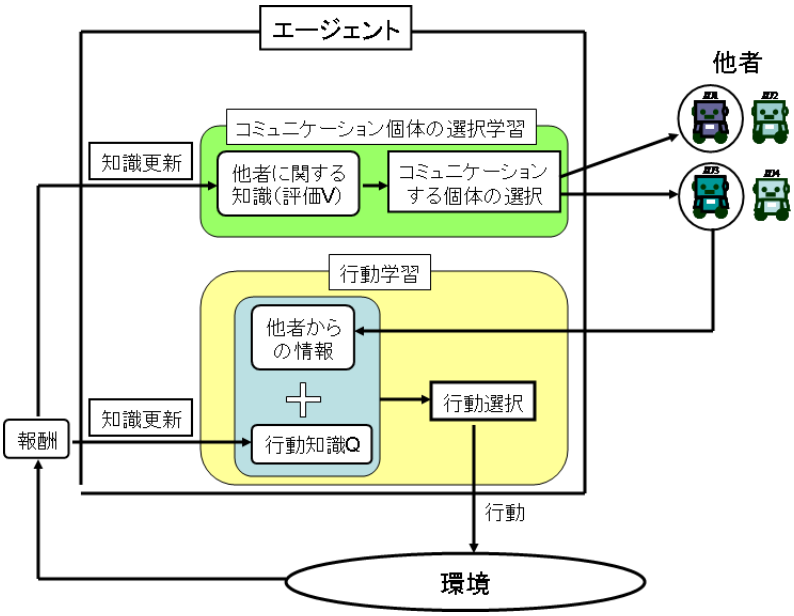

図 3.1: システムの概念図

囲ってある部分である．コミュニケーション個体の選択学習と行動学習の 2 つの部分でそれぞれ別な強化学習を使用する．これは，他者に関する知識と行動知識という 2 つの異なった知識を更新するためである．

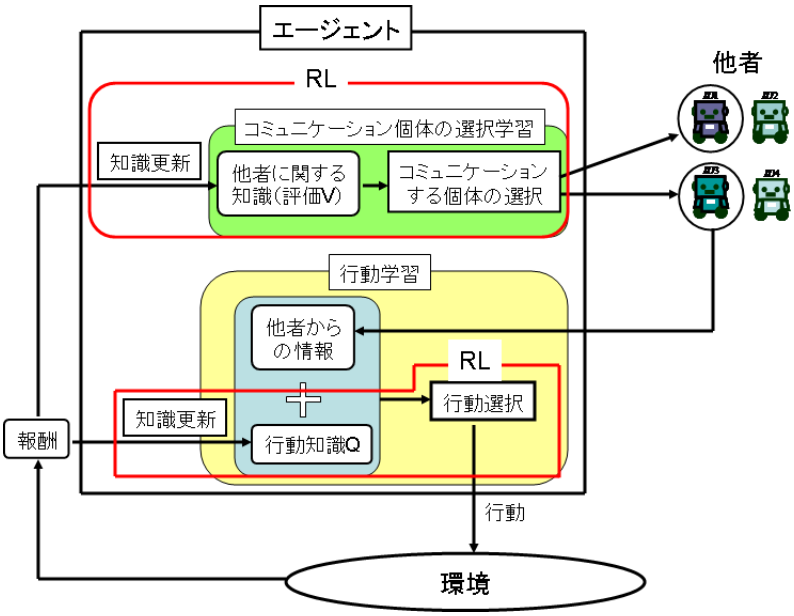

図 3.2: システムの概念図 RL 採用部

作成するシステムの流れを図 3.3 に示す．まず，他者に対する知識を基にコミュニケーションする個体を決定し，コミュニケーションする．次にコミュニケーションした情報と自身の知識から行動を選択する．そして，行動の結果得られた報酬から他者に対する知識と自身の行動知識を更新する．この流れを繰り返すことで学習を進める．

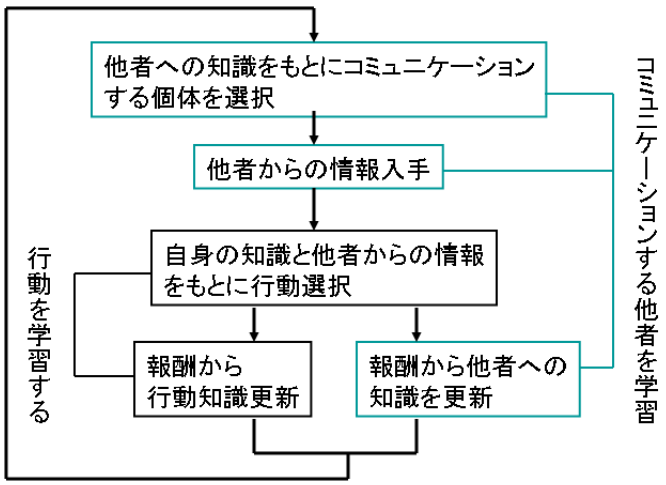

図 3.3: システムの流れ

3.2 コミュニケーションに用いる情報

本研究では強化学習の枠組みを用いて問題を考える．強化学習では，学習空間は，知覚能力 s と意思決定能力 a と知識 $Q(s,a)$ が構成する空間となる (図 3.4)．エージェントはこの学習空間を探索し，知識 $Q(s,a)$ を更新していく．コミュニケーションに用いる情報としては，自身から送信する情報として，現在直面している状態 s_k を送る．そして，他者からは状態 s_k において，選択することの出来るすべての行動に対する Q 値を得る（図 3.5）．

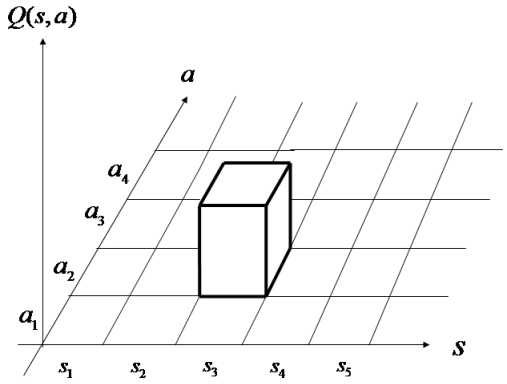

図 3.4: エージェントの学習空間

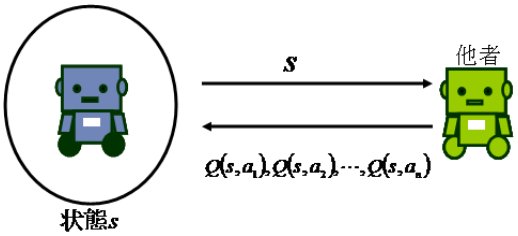

図 3.5: コミュニケーション情報

3.3 コミュニケーションする個体の取捨選択

コミュニケーションする他者の選択は，自身の他者に対する評価を元にして行なう．ある他者への評価が高い場合，その他者から提供される情報は自身にとって有益であると考えられる．コミュニケーションはそのような他者とコミュニケーションを行なうことが望ましい．そこで，他者への評価が高いほど高確率でコミュニケーションを行なうようにする．個体 i の保持する評価値は図??のようにエージェントごとに保持する．この評価値はエージェントが直面する状態ごとに保持するものではなく，環境に対して 1 つだけ保持する．つまり，状態に依存しない知識となる．

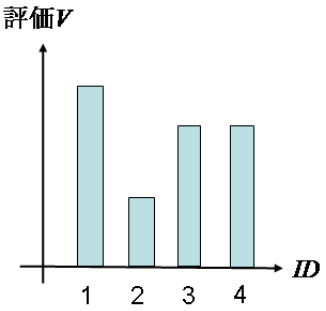

図 3.6: 自身の他者に対する知識

$V_i(j)$ を個体 i の個体 j に対する評価値とすると，個体 i が個体 j をコミュニケーション相手として選択する確率 $P_i(j)$ は式 (3.2) で算出される．式 (3.1) で自分自身の評価も含め，すべての個体の評価値を比較し最大のものを V_max として算出する．そして，式 (3.2) では，自分以外の他者に対してそれぞれ V_{max} を基準として選択確率を算出する．自分自身に対する評価も含め最大評価値 V_{max} を算出することで，他者を信じきることが無いようにする．特

に学習初期では，他者とコミュニケーションするよりも，自分自身の経験を頼りに行動したほうがよい場合が多いと考えられる．このようなときは，コミュニケーションを控える必要がある．そのため，自分自身も評価基準とするように V_{max} を算出する．

$$V_max = max_j V_i(j) \tag{3.1}$$

$$P_i(j) = \frac{V_i(j)}{V_{max}} \quad (i \neq j) \tag{3.2}$$

3.4 他者に対する評価の更新方法

他者に対する評価の更新方法は，入手した情報と実際に入手した報酬との差を基にして決定する．他者から提供される情報とその情報を基に自身が実際に入手した報酬の差が大きければ評価を下げ，小さければ評価を上げる．これは，情報と実際の結果が近ければ近いほど自身にとってその他者の情報は正しく，それに伴いその他者自体も信用できるという考え方を基にしている．今回，他者から提供される情報は，自身が直面している状態に対して，その状態でとることのできるすべての行動の Q 値である．しかし，自身が選択し結果を得ることが出来るのは，1 行動のみなので，評価はその行動のみに絞り評価する．

状態行動対 (s_k, a_l) に対する個体 i の入手した報酬 r_i と個体 j の提供する情報の差 $D_i(j)$ は式 (3.3) で算出される．

$$D_i(j) = |r_i - Q_j(s_k, a_l)| \tag{3.3}$$

個体 i から個体 j への評価を $V_i(j)$ とすると，評価の更新は $D_{i,j}$ を基に式 (3.4) で行なわれる．τ_v は学習率 $(0 \leq \tau_v \leq 1)$ である．なお，この信頼度の更新は情報交換を行なった全ての他者に対して行なわれる．

$$V_i(j) \leftarrow V_i(j) + \left(e^{-\frac{Diff}{\tau_v}} - V_i(j)\right) \tag{3.4}$$

3.5 コミュニケーション情報の利用

行動の選択は自身の Q 空間と他者の情報を合成して一時的な Q 空間を作成し，その空間を用いて，行動選択を行なう（図 4.8）．したがって，自身の Q 値空間が他者の Q 空間によって改変されることはない．このため，他者情報が自身の知識を改変することによる学習効率の低下が起きない．

状態行動対 (s_k, a_l) において，個体 i の Q 値空間を $Q_i(s_k, a_l)$，他者 j の Q 値空間を $Q_j(s_k, a_l)$ とすると一時的に作成する個体 i の Q 値空間 Q_{tmp_i} は式 (4.8) によって定義する．なお，γ_{tmpQ} は割引率 $(0 \leq \gamma_{tmpQ} \leq 1)$ である．式 (4.8) は，状態 s_k でとることの出来る全ての行動に対して実行される．

$$Q_{tmp_i}(s_k, a_l) = Q_i(s_k, a_l) + \gamma_{tmpQ} \sum_{j \in M} Q_j(s_k, a_l) \tag{3.5}$$

(M はコミュニケーションを行なった個体の集合)

3.6 行動学習

行動学習手法として加重平均手法を用いる．個体 i の現在の状態を s_k，そのときとった行動を a_l とし，このときの Q 値を $Q_i(s_k, a_l)$ とすると，Q 値の更新は式 (4.8) によって行なう．

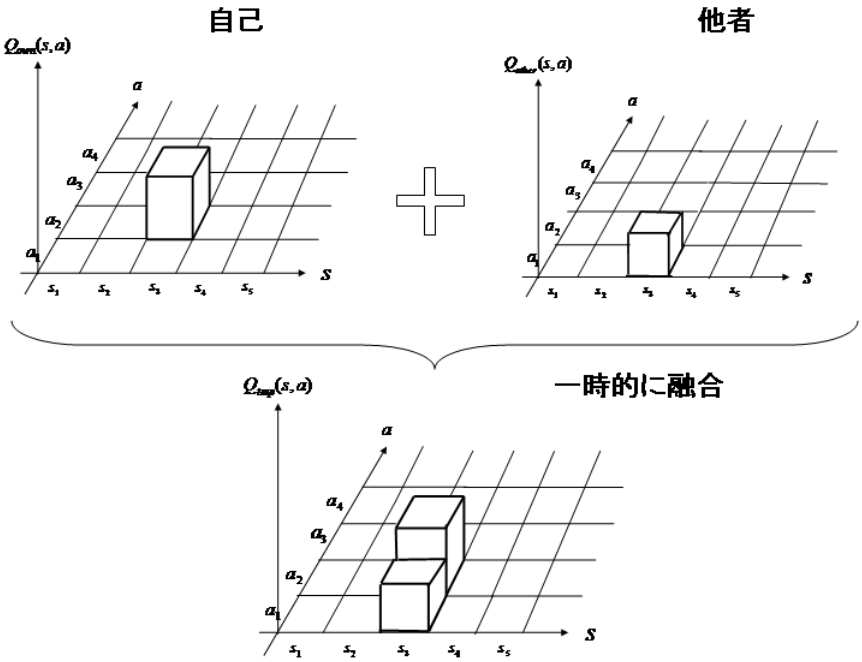

図 3.7: 一時的な Q 値空間の作成（この例では他者は 1 体だが，実際には複数いる可能性がある）

r は報酬，α_{act} は学習率 $(0 \leq \alpha_{act} \leq 1)$ である．また行動選択手法としては softmax 法を使用する．

$$Q_i(s_k, a_l) \leftarrow Q_i(s_k, a_l) + \alpha_{act}[r - Q_i(s_k, a_l)] \tag{3.6}$$

3.7 実験：迷路問題への適用

3.7.1 実験概要

提案システムの有効性の検証のために迷路タスクに適用する．エージェントはゴールまでの行動数を最小にするように学習する．本実験で用いる迷路タスクでは，各エージェントに対し数個のスタート・ゴールからそれぞれランダムに割り当てる．迷路の構造としては，ゴールまでのルートは 1 つではなく複数あるものを考える．また，コミュニケーションは知識量が多いものと知識量が少ないものとが行なう場合が最も効果的に作用すると考えられる．そこで，本実験ではエージェントが一定ステップ数ごとに上限数に達するまで個体の追加を行い，上限数に達した場合は古いエージェントから順に新しいエージェントと交換を行なう．これにより新しい個体（知識量が少ない）と古い個体（知識量が多い）が共存する環境になる．実験はエージェントのステップ数が上限に達するまで行なう．

結果は提案システムとランダムにコミュニケーションを行なったもの，全ての個体とコミュニケーションを行なったもの，コミュニケーションを行なわないものの 4 つの手法で比較を行なう．

3.7.2　本実験で作成する迷路環境の意義と作成方法

本実験で作成する迷路環境の意義

本実験で作成する迷路は，ゴールに至る道のりが複数種類ある迷路を考える．複数の解を用意することでエージェント個々の解に多様性が生まれ，解に優劣が存在する．よってコミュニケーションを通してより優れた解を獲得可能な環境である．このような複数の解が存在する環境では，単体学習では局所解に陥りやすい．コミュニケーションによる他者と情報交換をすることで局所解を脱出しよりよい解へたどり着くことが可能である．また，各エージェントに複数のスタート・ゴールを割り当てることで，同一環境上でも異なった状況（スタート・ゴール）におかれるようにした．これにより様々な状況（スタート・ゴール）におかれる個体が存在する．そのため，自身と似た状況に置かれる個体とコミュニケーションしないと有益な情報が得られない．コミュニケーション相手を適切に選択することが高い報酬につながるような迷路環境を作成することで，提案手法の優位性をより顕著に確認することが出来ると考えられる．

迷路環境の作成方法

迷路環境の作成は棒倒し法を用いて行なう．棒倒し法は，図 4.9 のように初期状態を生成する．次に図 4.10 のように，最初の 1 行は上下左右にランダムに壁を作る．そして 2 行目以降は上方向以外の下左右に対しランダムで壁を作っていく．すでに壁があった場合はそのまま次のマスに移る．このようにすることでゴールまでのルートが複数存在する迷路を作成することができる．棒倒し法が終了した時点の例を図 4.11 に示す．この後図 4.11 に複数のスタート・ゴールを設定することで迷路が完成する．本実験では，スタート・ゴールはそれぞれ一定数設置する．設置方法は，スタートの場合は 1 行目の通路上にランダムに配置する．ゴールの場合は最終行の通路上にランダムに配置する．スタートを 2 箇所，ゴールを 3 箇所配置した場合の例を図 4.12 に示す．エージェントはそれぞれスタートとゴールをランダムに指定される．エージェントは指定されたスタート・ゴール間で最も報酬を得られるルートを探索する．

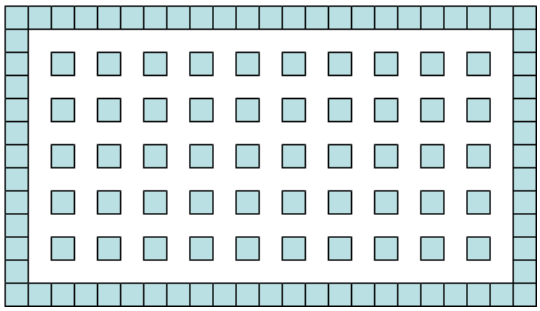

図 3.8: 棒倒し法初期状態

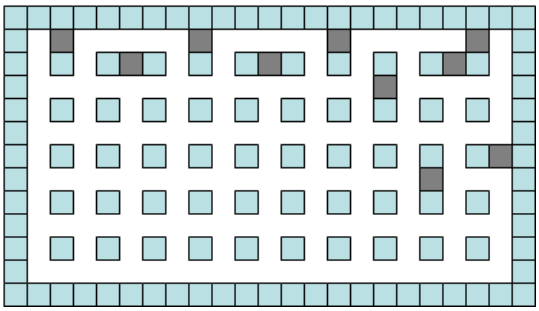

図 3.9: 棒倒し法イメージ

3.7.3　パラメータ設定

迷路に関する設定を表 1 に，エージェントに関する設定を表 4.1 に示す．

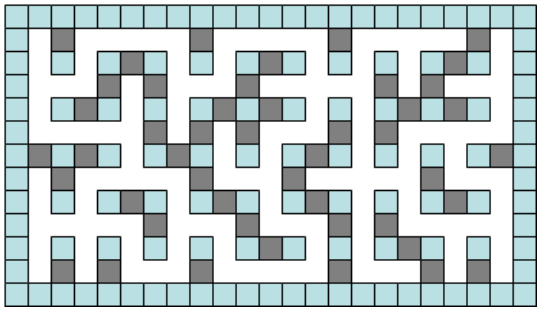

図 3.10: 棒倒し法終了時の迷路

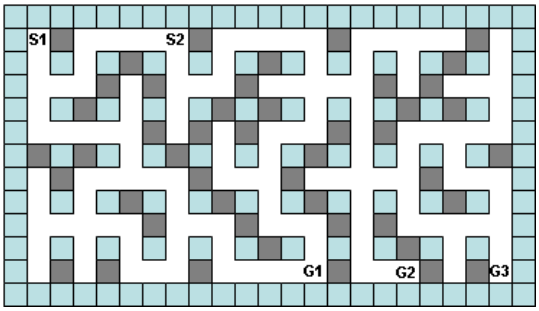

図 3.11: スタート・ゴールの設定の例

表 3.1: パラメータ設定

迷路のサイズ	50×50
スタートの数	6
ゴールの数	6
ゴール報酬	20
エージェントの数	80
総ステップ数	80000
循環ステップ数	600
τ_v	4
α_{act}	0.1
γ_{tmpQ}	0.05

3.7.4 実験結果・考察

実験結果を図 3.12 に示す．図 3.12 は各ステップの 1 個体あたりの平均獲得報酬量である．ステップ数が 48000 回まではエージェント数は上限数（80 体）存在せず，循環ステップ数（600 ステップ）ごとに 1 体ずつ追加されている．48600 ステップ以降は循環ステップ数ごとに最も古い個体を取り除き新しい個体を加えている．個体の循環が行なわれるようになるのは 48600 ステップ以降である．図 3.12 では 10000 ステップあたりから提案手法の平均獲得報酬量が他の手法を上回っている．10000 回以降からはエージェント数は 16 体程度となり，共通するスタート，またはゴールを持つ個体も徐々に増加してきている．そのため，コミュニケーション相手の取捨選択が有効に作用し始め，獲得報酬の増加につながっていると考えられる．以降エージェント数が増えていくと共に獲得報酬量も増加していく．これはエージェントが増えることで自身と共通のスタート・ゴールを持つ個体が増え，よりコミュニケーション相手の取捨選択が有効に働くようになったためである．

全体とコミュニケーションした場合は試行 60000 回以降を除き基本的にはランダムの場合より高い．これは，全体とコミュニケーションすることで，有益な情報と不利益な情報の両方が入ってくる．そのため不利益な情報に自身の意思決定が影響され，提案手法ほど高い獲得報酬は得られなかったと考えられる．

ランダムにコミュニケーションを行なった場合では 50000 ステップ程度までは全体とコミュニケーションした場合に劣る平均獲得報酬量であった．今回は現在のエージェント数を超えない範囲でランダム個体数，ランダムに対象を選択した．そのため，自身に良い影響を与える情報は入っていくるが，同時に悪い情報も同程度に入ってくる．入ってくる情報によって自身にとって良い行動を行うことと悪い行動を行うことが傾向がなく起こる．これにより，自身にとっていい個体とコミュニケーションを行ってもその後すぐに自身に悪影響を

与える個体とコミュニケーションしてしまうことが起きやすい．したがって，あまり獲得報酬量としては少なかったと考えられる．

総じて，コミュニケーションを行なった方が高い平均獲得報酬量であることからコミュニケーションは個体知能発達に有効である．また，提案手法が最も良好な結果である．以上から，即時報酬環境下における提案手法の有効性を検証することが出来た．また，コミュニケーション相手の取捨選択がコミュニケーションを用いた個体知能発達を促進していることがわかった．

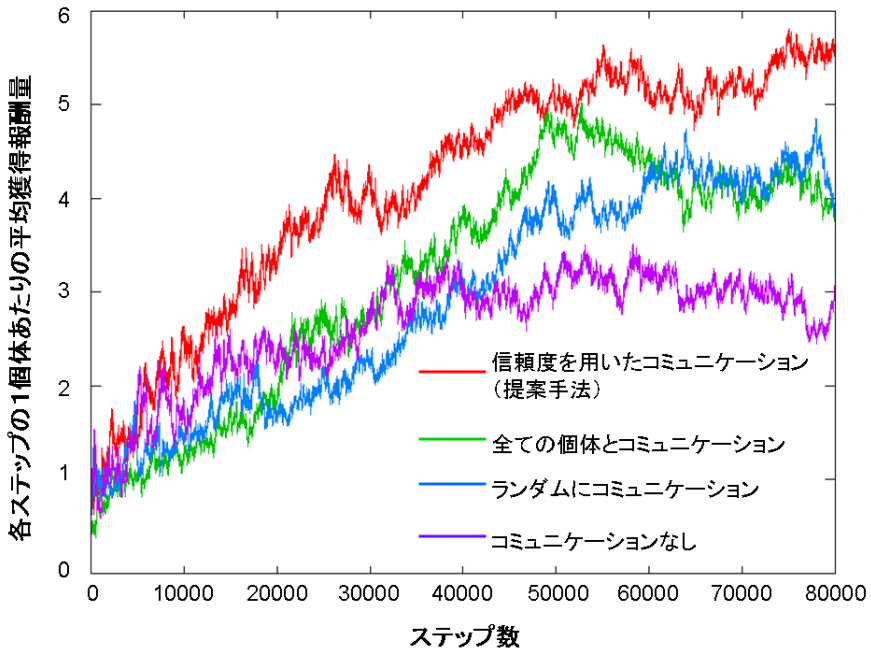

図 3.12: 各ステップの 1 個体あたりの平均獲得報酬量

3.8 まとめ

本章では，即時報酬環境でのコミュニケーション相手の取捨選択に関する手法を提案した．また，遅延報酬迷路環境に提案手法を適用し提案手法の有効性を示した．

第4章　遅延報酬環境下でのコミュニケーション相手の選択による効率的知能発達

本章では，2 章で紹介した強化学習の枠組みを用いてシステムを構築し，実験により有効性を検証する．強化学習は実ロボットに使われることが多い手法である [28]-[32]．本システムもいづれは実ロボットへの適用を考えているので本手法では強化学習を用いてシステムを構築する．本章で構築するシステムは，行動の度に報酬が与えられるのではなく，目的た達成されたときにのみ報酬が入手できるような遅延報酬型の環境に対しての手法である．

4.1　作成するシステムの概要

作成するシステムの概念図を図 4.1 に示す．本システムは，コミュニケーション個体の選択を学習する部分と直面する状況に対して適切な行動を学習する部分の 2 つの学習を行う．エージェントが保持する知識は，他者に関する知識（自身から他者に対しての評価 (V)）と行動に関する知識 (Q) となる．他者に関する知識 V はコミュニケーション相手の取捨選択の際の指標なり，行動に関する知識 (Q) は直面する状況に対して行動を選択する際の指標となる．それぞれの知識は，行動の結果得られる報酬を基にして更新，蓄積される．他者からの情報は，自身の知識とともに行動選択に利用することで自身の保持する知識量を増やし，行動選択に反映する．エージェントは自身にとって適切なコミュニケーション相手を学習すると同時に直面する状況に適した行動を学習する． 強化学習の適用部は図 3.2 の赤い枠で

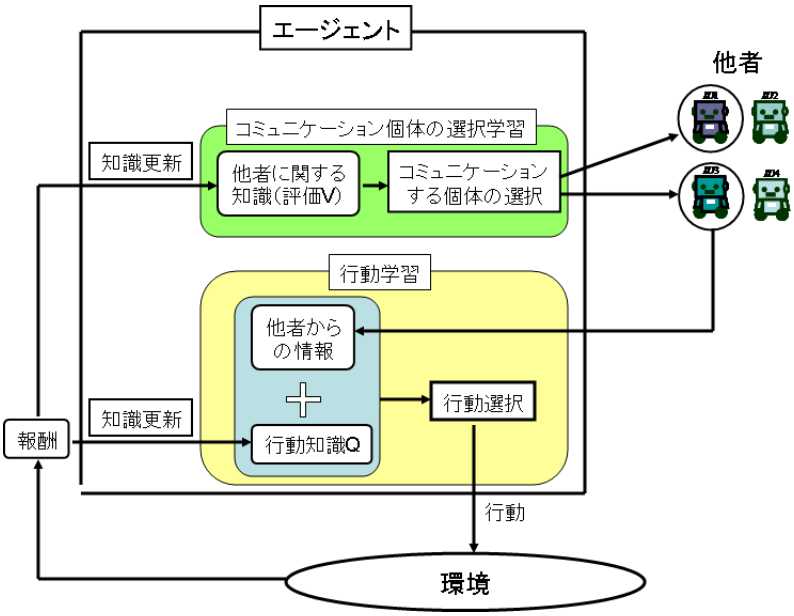

図 4.1: システムの概念図

囲ってある部分である．コミュニケーション個体の選択学習と行動学習の 2 つの部分でそれぞれ別な強化学習を使用する．これは，他者に関する知識と行動知識という 2 つの異なった知識を更新するためである．

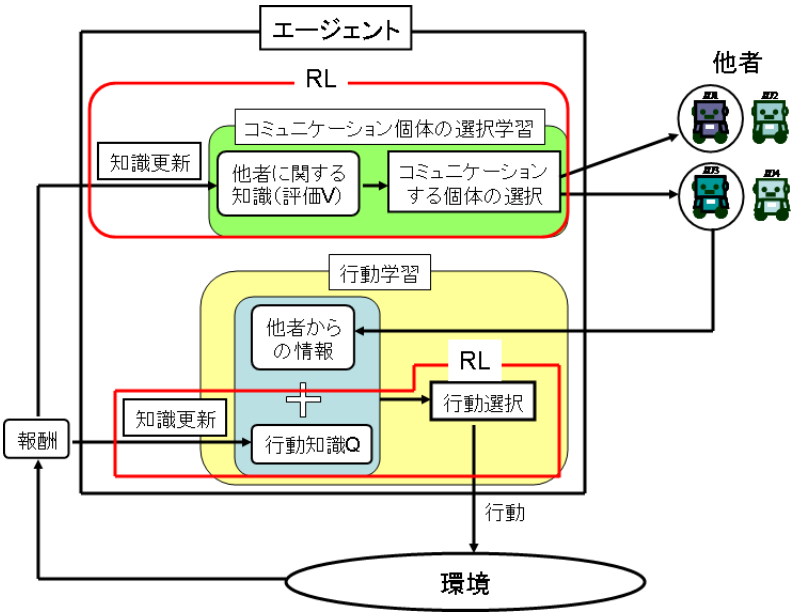

図 4.2: システムの概念図 RL 採用部

作成するシステムの流れを図 3.3 に示す．まず，他者に対する知識を基にコミュニケーションする個体を決定し，コミュニケーションする．次にコミュニケーションした情報と自身の知識から行動を選択する．そして，行動の結果得られた報酬から他者に対する知識と自身の行動知識を更新する．この流れを繰り返すことで学習を進める．

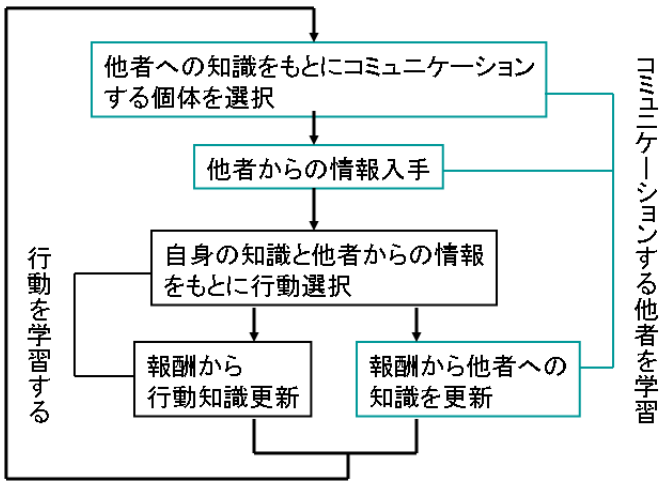

図 4.3: システムの流れ

4.2 コミュニケーションに用いる情報

コミュニケーションに用いる情報について述べる．本手法の性質上，情報の請求元（自身）は現在の状態 s を相手に送信する．相手は，その状態において最も良いとされる行動 a_{best} とその評価値 $Q(s, a_{best})$ とする（図 4.4）．

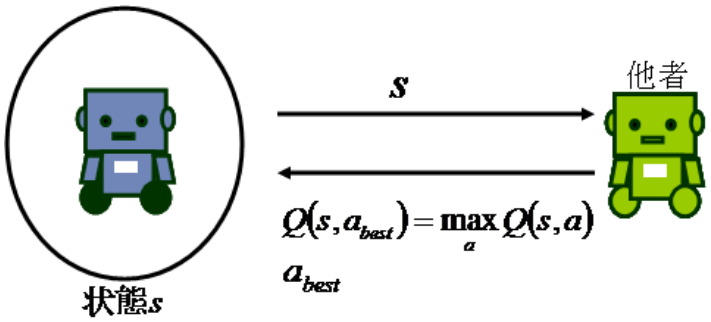

図 4.4: コミュニケーション情報

4.3 コミュニケーションする個体の取捨選択※

コミュニケーションする他者の選択は，自身の他者に対する評価を元にして行なう．ある他者への評価が高い場合，その他者から提供される情報は自身にとって有益であると考えられる．コミュニケーションはそのような他者とコミュニケーションを行なうことが望ましい．そこで，他者への評価が高いほど高確率でコミュニケーションを行なうようにする．個体 i の保持する評価値は図??のようにエージェントごとに保持する．この評価値はエージェントが直面する状態ごとに保持するものではなく，環境に対して 1 つだけ保持する．つまり，状態に依存しない知識となる．

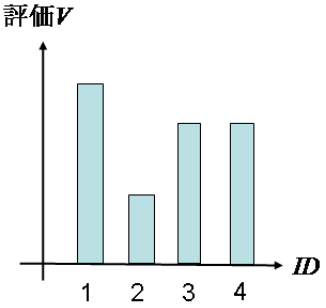

図 4.5: 自身の他者に対する知識

$V_i(j)$ を個体 i の個体 j に対する評価値とすると，個体 i が個体 j をコミュニケーション相手として選択する確率 $P_i(j)$ は式 (3.2) で算出される．式 (4.1) で自分自身の評価も含め，すべての個体の評価値を比較し最大のものを V_max として算出する．そして，式 (4.2) では，自分以外の他者に対してそれぞれ V_{max} を基準として選択確率を算出する．自分自身に対する評価も含め最大評価値 V_{max} を算出することで，他者を信じきることが無いようにする．特に学習初期では，他者とコミュニケーションするよりも，自分自身の経験を頼りに行動したほうがよい場合が多いと考えられる．このようなときは，コミュニケーションを控える必要がある．そのため，自分自身も評価基準とするように V_{max} を算出する．

$$V_max = max_j V_i(j) \tag{4.1}$$

$$P_i(j) = \frac{V_i(j)}{V_{max}} \quad (i \neq j) \tag{4.2}$$

4.4 他者に対する評価の更新方法

他者に対する評価の更新方法について説明する．即時報酬環境（3 章）では，エージェントが行動する度に受け取った報酬から他者評価を更新していた．しかし，遅延報酬環境では目的を達成したときにのみ報酬が与えられるという性質上，エージェントが行動する度に他者評価をすることができない．そこで，報酬を受け取った時点で過去に遡って他者評価を行う．評価はエージェントが報酬を受け取るまでの各ステップごとに受け取った報酬を基にして，評価対象となる他者に対して行う．評価対象となる他者は，エージェントがコミュニケーションを行い情報を採用された他者となる．本手法においてコミュニケーション情報は，自身の現在状態 s における最良行動 a_{best} とその評価値 $Q(s, a_{best})$ である．したがって，エージェントの選択した行動と同じ行動を情報として提供した他者が，情報を採用された他者となる．他者情報の採用の概念図を図 4.6 にしめす．図 4.6 のように各時間ごとに情報を採用した個体を評価対象として記録しておく． この評価対象を記録を個体毎に定式化したもの

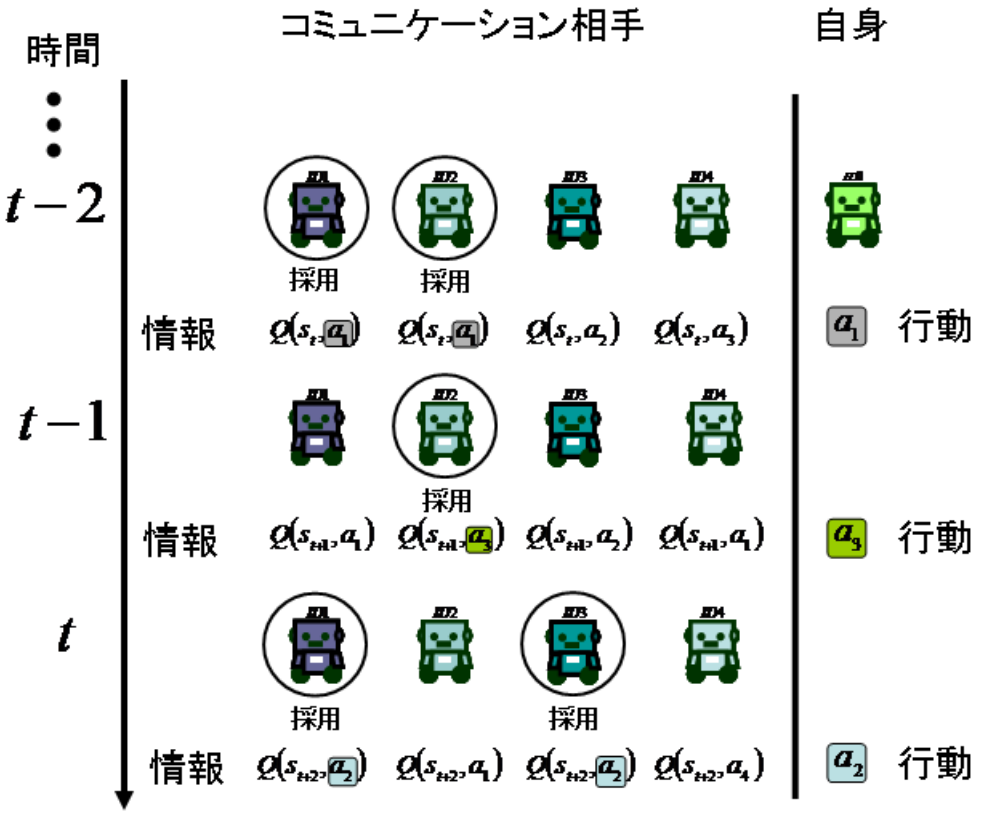

図 4.6: 情報を採用された他者の例

が式 4.3 である．$CommLog_i(j)$ は個体 i の個体 j に対する情報採用の記録である．式 4.3 に示す．報酬を受け取った時のステップ数を t とする．$CommLog_i(j)$ の Log_t には t 時点までの各時点で情報を採用したかどうかが格納される．格納される値は式 4.4 に従う．情報を採用した場合には 1 が，情報が採用されなかったまたはコミュニケーションを行っていない場合には 0 が格納される．記録の格納イメージを図 4.7 に示す．図 4.7 のようにエージェントはすべての他者に対して採用したかどうかの情報を保持する．

$$CommLog_i(j) = [\ Log_1 \quad Log_2 \quad \cdots \quad Log_t\ ] \tag{4.3}$$

$$Log_t = \begin{cases} 1 & (\text{コミュニケーションを行いかつ情報を採用した個体}) \\ 0 & (\text{コミュニケーションしていない，または情報を採用していない場合}) \end{cases} \tag{4.4}$$

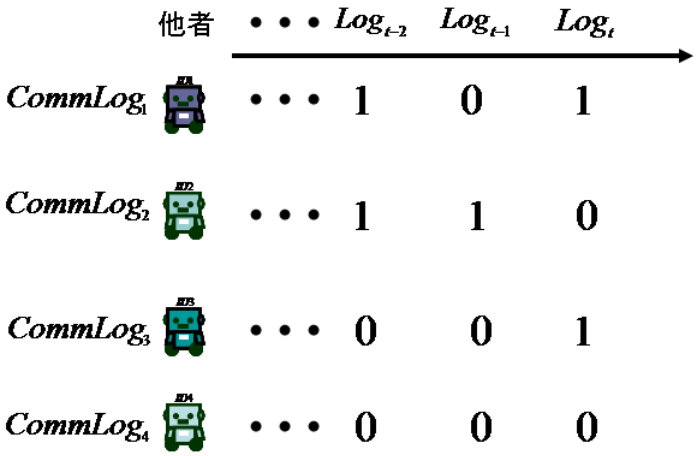

図 4.7: CommLog の格納イメージ

他者の評価は報酬を入手した段階で行なわれる（迷路タスクならばゴールした時点）．評価の基となる情報は報酬と $CommLog$ である．また，評価はより最近にコミュニケーションした他者に対して大きい重みをつける．これは，より最近の情報ほど報酬獲得に貢献する情報を提供した個体である可能性が高いためである．更新式を式**??**に示す．$V_i(j)$ は個体 i の個体 j に対する評価値を表す．$CommLog_i(j)$ は個体 i の個体 j に対する評価値を表す．また，r は報酬，Γ は割引率の累乗の行列である．行列 Γ は割引率 γ の累乗が行列の要素として格納されており，最後の列要素に近くなるにつれ重みが増すようになっている（式 4.6）．$A_i(j)$ は個体 i が個体 j の情報を採用した回数である．式**??**は 1 回のコミュニケーション当たりに得られる期待報酬を計算している．なお，γ は $0 \leq \gamma_v \leq 1$，α_v は $0 \leq \alpha_v \leq 1$ の範囲の値をとる．

$$V_i(j) \leftarrow V_i(j) + \alpha_v(fracr \times CommLog_i(j) \times \Gamma A_i(j) - V_i(j)) \tag{4.5}$$

$$\Gamma = [\ \gamma_v^{t-}1 \quad \gamma_v^{t-2} \quad \cdots \quad 1\ ] \tag{4.6}$$

4.5 コミュニケーション情報の利用

行動の選択は自身の Q 空間と他者の情報を合成して一時的な Q 空間を作成し，その空間を用いて，行動選択を行なう（図 4.8）．したがって，自身の Q 値空間が他者の Q 空間によって改変されることはない．このため，他者情報が自身の知識を改変することによる学習効率の低下が起きない．

状態行動対 (s_k, a_l) において，個体 i の Q 値空間を $Q_i(s_k, a_l)$，他者 j の Q 値空間を $Q_j(s_k, a_l)$ とすると一時的に作成する個体 i の Q 値空間 Q_{tmp_i} は式 (4.8) によって定義する．なお，γ_{tmpQ} は割引率 $(0 \leq \gamma_{tmpQ} \leq 1)$ である．式 (4.8) は，状態 s_k でとることの出来る全ての行動に対して実行される．

$$Q_{tmp_i}(s_k, a_l) = Q_i(s_k, a_l) + \gamma_{tmpQ} \sum_{j \in M} Q_j(s_k, a_l) \tag{4.7}$$

(M はコミュニケーションを行なった個体の集合)

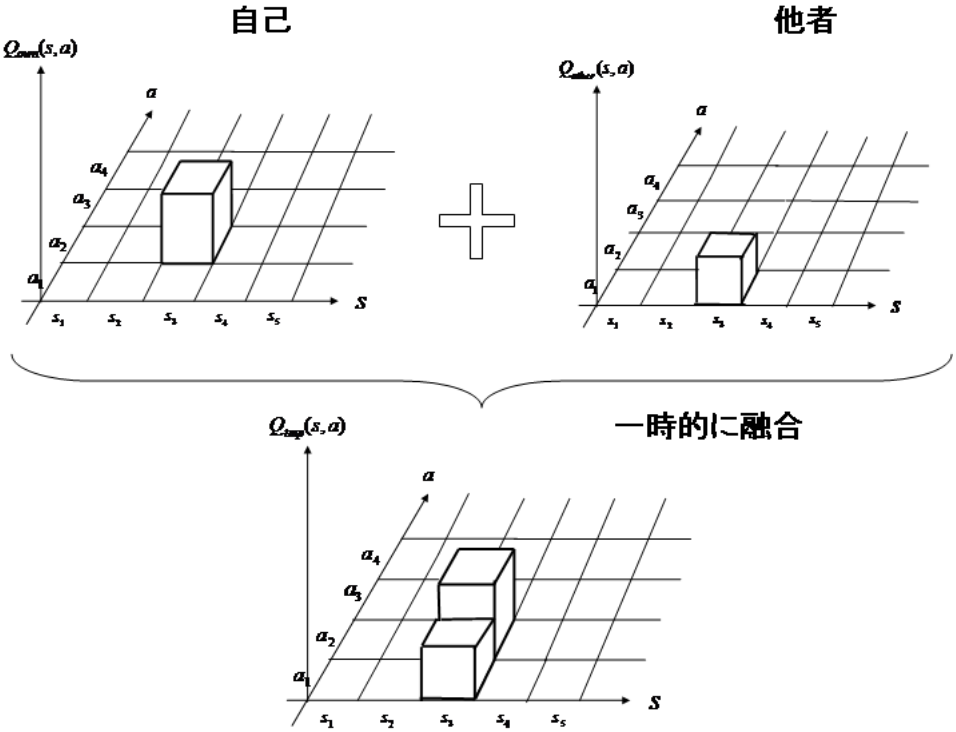

図 4.8: 一時的な Q 値空間の作成（この例では他者は 1 体だが，実際には複数いる可能性がある）

4.6 行動学習

行動学習手法として Q 学習を用いる．個体 i の現在の状態を s_k，そのときとった行動を a_l とし，このときの Q 値を $Q_i(s_k, a_l)$ とすると，Q 値の更新は式 (4.8) によって行なう．r は報酬，α_{act} は学習率 $(0 \leq \alpha_{act} \leq 1)$ である．また行動選択手法としては ϵ-greedy 法を使用する．

$$Q_i(s_k, a_l) \leftarrow Q_i(s_k, a_l) + \alpha_{act}[r - Q_i(s_k, a_l)] \tag{4.8}$$

4.7 実験：迷路問題への適用

4.7.1 実験概要※

提案システムの有効性の検証のために迷路タスクに適用する．エージェントはゴールまでの行動数を最小にするように学習する．本実験で用いる迷路タスクでは，各エージェントに対し数個のスタート・ゴールからそれぞれランダムに割り当てる．迷路の構造としては，ゴールまでのルートは 1 つではなく複数あるものを考える．また，コミュニケーションは知識量が多いものと知識量が少ないものとが行なう場合が最も効果的に作用すると考えられる．そこで，本実験ではエージェントが一定ステップ数ごとに上限数に達するまで個体の追加を行い，上限数に達した場合は古いエージェントから順に新しいエージェントと交換を行なう．これにより新しい個体（知識量が少ない）と古い個体（知識量が多い）が共存する環境になる．実験はエージェントのステップ数が上限に達するまで行なう．

結果は提案システムとランダムにコミュニケーションを行なったもの，全ての個体とコミュニケーションを行なったもの，コミュニケーションを行なわないものの 4 つの手法で比較を行なう．

4.7.2　本実験で作成する迷路環境の意義と作成方法

本実験で作成する迷路環境の意義

本実験で作成する迷路は，ゴールに至る道のりが複数種類ある迷路を考える．複数の解を用意することでエージェント個々の解に多様性が生まれ，解に優劣が存在する．よってコミュニケーションを通してより優れた解を獲得可能な環境である．このような複数の解が存在する環境では，単体学習では局所解に陥りやすい．コミュニケーションによる他者と情報交換をすることで局所解を脱出しよりよい解へたどり着くことが可能である．また，各エージェントに複数のスタート・ゴールを割り当てることで，同一環境上でも異なった状況（スタート・ゴール）におかれるようにした．これにより様々な状況（スタート・ゴール）におかれる個体が存在する．そのため，自身と似た状況に置かれる個体とコミュニケーションしないと有益な情報が得られない．コミュニケーション相手を適切に選択することが高い報酬につながるような迷路環境を作成することで，提案手法の優位性をより顕著に確認することが出来ると考えられる．

迷路環境の作成方法

迷路環境の作成は棒倒し法を用いて行なう．棒倒し法は，図 4.9 のように初期状態を生成する．次に図 4.10 のように，最初の 1 行は上下左右にランダムに壁を作る．そして 2 行目以降は上方向以外の下左右に対しランダムで壁を作っていく．すでに壁があった場合はそのまま次のマスに移る．このようにすることでゴールまでのルートが複数存在する迷路を作成することができる．棒倒し法が終了した時点の例を図 4.11 に示す．この後図 4.11 に複数のスタート・ゴールを設定することで迷路が完成する．本実験では，スタート・ゴールはそれぞれ一定数設置する．設置方法は，スタートの場合は 1 行目の通路上にランダムに配置する．ゴールの場合は最終行の通路上にランダムに配置する．スタートを 2 箇所，ゴールを 3 箇所配置した場合の例を図 4.12 に示す．エージェントはそれぞれスタートとゴールをランダムに指定される．エージェントは指定されたスタート・ゴール間で最も報酬を得られるルートを探索する．

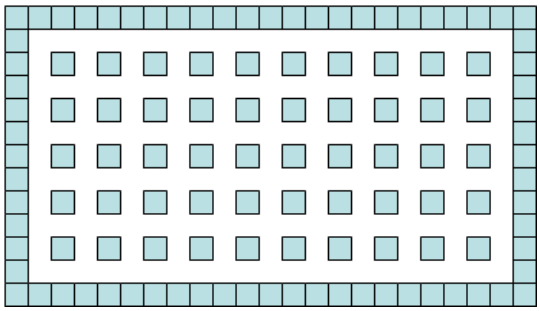

図 4.9: 棒倒し法初期状態

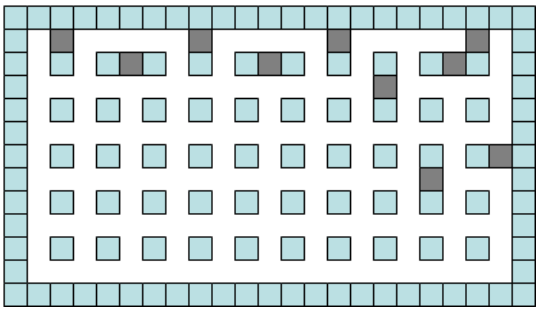

図 4.10: 棒倒し法イメージ

4.7.3　パラメータ設定

迷路に関する設定を表 1 に，エージェントに関する設定を表 4.1 に示す．

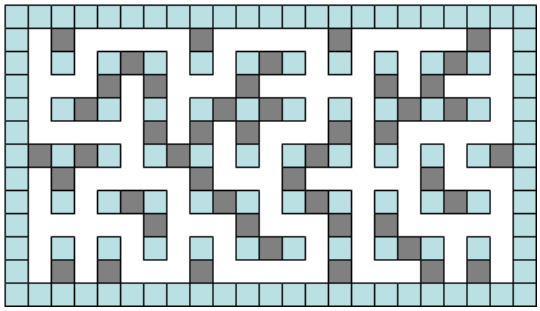

図 4.11: 棒倒し法終了時の迷路

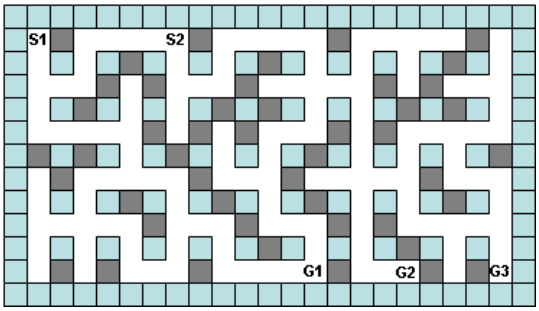

図 4.12: スタート・ゴールの設定の例

表 4.1: パラメータ設定

迷路のサイズ	21×21
スタートの数	1
ゴールの数	3
ゴール報酬	1
エージェントの数	50
総ステップ数	600000
循環ステップ数	3000
α_v 0.5	
γ_v	0.9
α_{act}	0.5
γ_{act}	0.8
γ_{tmpQ}	0.01
ϵ	0.05

4.7.4 実験結果・考察

実験結果を図 4.13 に示す．図 4.13 から，コミュニケーション相手の取捨選択を行った場合の 1 個体あたりの生涯獲得報酬量は他の手法に比べ多いことがわかる．また図 4.14 から，群全体で見た場合の生涯獲得報酬量の総和についてもコミュニケーション相手の取捨選択を行った方がより多くの報酬を獲得していることがわかる．これは，コミュニケーション相手の取捨選択により，行為主体者は自身の発達に有益な情報を多く入手することができる．そのため，コミュニケーション相手の取捨選択が行動学習の効率を向上したことがいえる．

ランダムにコミュニケーションを行った場合と全個体とコミュニケーションを行った場合では，世代数が 30〜50 世代のあたりでコミュニケーションなしの手法に生涯獲得報酬量において低い結果となっている．これは，ランダムにコミュニケーションを行う場合と全ての個体とコミュニケーションを行う場合では，悪影響を与える情報を取り入れやすいためである．ランダムにコミュニケーションする場合では，コミュニケーション個体数と相手をランダムで決定するため自身に悪影響を与える情報を持つ個体ともコミュニケーションを行ってしまう．

以上のことから，自身にとって有益な他者を学習するまでは一時的に学習効率は落ちるが，他者の学習が完了すると行動学習の効率が上昇するといえる．したがって，提案手法はごく短時間の学習時間では効率的な学習は行えないが，長期的な学習では十分に効率的な学習が

可能になるといえる．

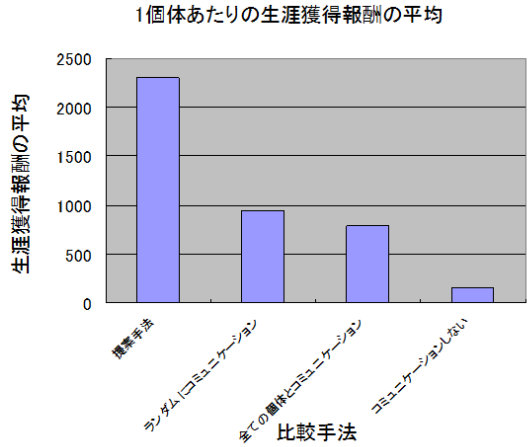

図 4.13: 1 個体あたりの生涯獲得報酬平均

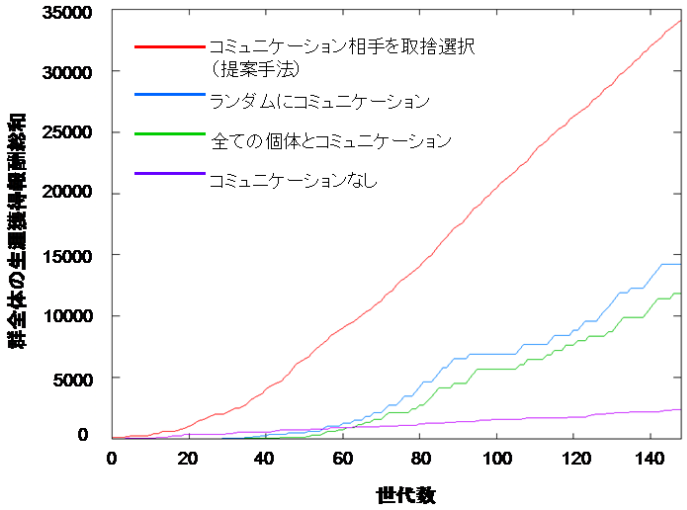

図 4.14: 群全体の獲得報酬総和

4.8 まとめ

本章では，遅延報酬環境でのコミュニケーション相手の取捨選択に関する手法を提案した．また，遅延報酬迷路環境に提案手法を適用し提案手法の有効性を示した．

第5章　結論

5.1　まとめ

コミュニケーション相手の選択により個体知能を効率的に発達させるシステムの構築を目的とした．本論文では，コミュニケーションを通して自身にとって有益な情報を提供する他者を取捨選択する．エージェントはコミュニケーションを通して他者の情報を入手する．入手した情報を基に行動した結果によって，情報を提供した他者を評価する．コミュニケーションした結果，良い結果に結びつけば，その結果に結びついた情報を提供した個体は自身にとって有益な情報を提供する個体となる．このような個体に対しては高く評価し，今後積極的にコミュニケーションを行う．逆にコミュニケーションケーションの結果，悪い結果結びつけばその結果に結びついた情報を提供した個体は自身にとって害となる情報を提供する個体となる．したがって，このような個体に対しては低い評価をつけ，以後コミュニケーションを控えるようにする．コミュニケーション個体の取捨選択は，実際にコミュニケーションにより他者から情報を入手する．そしてそれを用いて行動した結果から他者を評価する．これを繰り返すことで，有益な情報を提供する個体を学習する．そして，有益な情報を持つ個体に絞りコミュニケーションすることで，自身に有害な情報を取り込む事は無くなる．ゆえに行動学習に使用する他者情報は自身に有益なものとなり，より効率的な学習を実現することが可能となる．

本論文では以上のような考えのもと，強化学習を用いてコミュニケーション個体を取捨選択することで個体学習を促進させるシステムを作成した．本システムは，2つの学習部分によって構成される（図 5.1）．一方は，状態に適した行動を学習する部位である．もう一方は，自身にとって有益な情報を提供するコミュニケーション個体を学習する部位である．これら2つの学習部に対して強化学習を用いてシステムを構築した．今回，適用する環境とし

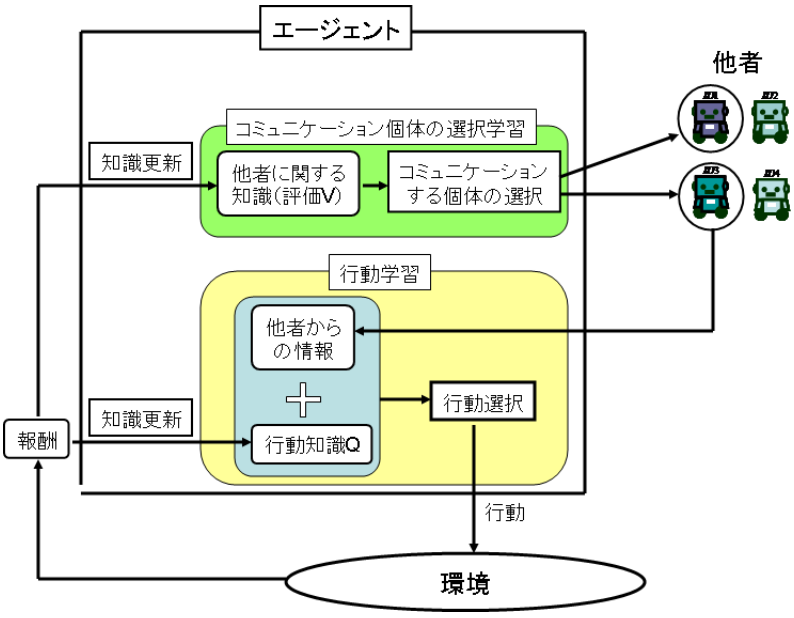

図 5.1: システムの概念図

て，即時報酬環境と遅延報酬環境の 2 種類の環境に適用する．それに際して，システムの概念は同じであるが，個々の学習部位で使用する学習手法はそれぞれ異なったものを使用する．よって，提案システムとしては，即時報酬環境に対応するものと遅延報酬に対応するものの 2 つのシステムを構築した．

検証実験として，即時報酬環境迷路問題と遅延報酬環境迷路問題に適用した．検証は提案システム・ランダムにコミュニケーションを行う場合・全個体とコミュニケーションする場合・コミュニケーションを行わない場合の 4 つの場合で比較を行った．即時報酬環境・遅延報酬環境ともに，学習初期は有効な結果は得られないが，学習が進むにつれて提案手法が他の手法に比べ優れた結果を示した．このことから提案システム有効性を検証することができた．

5.2 今後の課題

5.2.1 他タスクでの検証

今回は迷路問題に手法を適用したが，もっとも他の種類のタスクに対しても適用したい．特に現実に即したタスクを考案しそれに対して本システムを適用し有効性の検証を行っていきたい．現在コンピュータシミュレーション上で有効であろうと思われるものとしては，ロボットサッカー [] やロボットレスキュー [?] がある．これらは，いずれも群を構成する個体それぞれに目的がある．特にロボットレスキューは災害救助という現実に即したタスクとなっているため，本システムの有効性の検証には適していると考えられる．

5.2.2 より良い他者の評価の仕方の考察

自身の経験から評価対象を決定

今回の遅延報酬環境での他者評価はゴールに達までの過去 t ステップ分を遡り，情報を採用した他者に対して評価を行う．この方法では，他者の情報が有益なものでなくても情報を採用した個体であれば評価されてしまう．このため，情報が採用される回数が多ければ多いほど評価される回数も多くなる．そのため，自身と目的が異なる個体のように自身の学習に悪影響を与える個体でも採用価数が多ければそれだけ評価も高くなり，結果としてコミュニケーション相手として選択される確率が高くなる．特に，学習初期においては，報酬を得るまでは適当にコミュニケーションをとり続ける．そのため自身の学習に悪影響を与える個体からの情報を採用し評価してしまう可能性が高い．この評価がそれ以降の学習にも影響してしまうため，効率的な学習を阻害してしまうと考えられる．

したがって今後の課題としては，自身の目的達成にどれだけ貢献したかを自身の経験から評価するような手法を考える．例えば，迷路問題の場合にはゴールまでの過去辿ってきた状態を記憶し，その中で最短ルートにつながる情報を提供した個体に関しては高い評価を与え，それ以外の個体には低い評価を与えるということが考えられる．これは，目的を達成したあとにゴールの決め手となる情報を提供した個体はどの個体だったかを特定する作業にあたる．

又聞きの情報に対する評価

人間が他者から情報を入手し利用する最に，信頼できる他者の情報はもちろんのこと，その他者が推薦する他者の情報も信頼するといったことがある．自身が信頼している他者が信頼する他者であるので，この 3 者はそれぞれ似たような目的や身体構造を持っている可能性が高い．したがって他者が推薦する他者からの情報も自身にとって有益なものである可能性

は高い．そこで，コミュニケーションの際に他者から提供される情報にその他者が最も信頼する他者の情報も提供する．これによって他者の評価に加えて他者が推薦する他者も評価することで，自身にとって有益なコミュニケーション相手の学習が効率的に行われることが期待される．

5.2.3 実ロボットに適用

本論文では，コンピュータシミュレーション環境でのみの手法の検証を行った．次は実ロボットを用いて手法の有効性の検証を行いたい．具体的なタスクとしては，迷路問題やロボットサッカー・ロボットレスキューといったことを考えいている．

参考文献

[1] 神田崇行, 石黒浩, 小野哲雄, 今井倫太, 前田武志, 中津良平, ”研究用プラットホームとしての日常活動型ロボット”Robovie”の開発”, 電子情報通信学会論文誌, D-I, Vol.J85-D-I, No.4, pp.380-389, 2002.

[2] 柴田 崇徳, ”人の心を豊かにするメンタルコミットロボット”, 日本機械学會誌, Vol.109, No.1051, 2006.

[3] 川内直人, 古結義浩, 長島是, 大西献, 日浦亮太, ”ホームユースロボット ”wakamaru””, 三菱重工技報, Vol.40, No.5, 2003.

[4] 濱田彰一, 間野隆久, ”欧米における原子力防災ロボットの調査報告”, 日本ロボット学会誌, Vol.19, No.6, pp.678-684, 2001.

[5] 田所諭, 大須賀公一, 天野久徳, ”レスキューロボット”, 日本ロボット学会誌, Vol.19, No.6, pp.685-688, 2001.

[6] 亀川哲志，松野文俊, ”遠隔操作性を考慮した双頭ヘビ型レスキューロボット KOHGA の開発”, 日本ロボット学会誌, Vol.25, No.7, pp.1074-1081，2007.

[7] 稲葉典康, ”宇宙機運用への「AI」技術応用の期待”, 人工知能学会誌, Vol.21, No.1, pp.14-19, 2006.

[8] 久保田考, ”惑星別探査ローバ” 日本ロボット学会誌, Vol.21, No.5, pp.468-471, 2003.

[9] 茂原正道, 西田信一郎, ”宇宙探査ローバの作り方”, 日本ロボット学会誌, Vol.21, No.5, pp.472-476, 2003.

[10] 金森洋史, ”月・惑星探査のテラメカニクス”, 日本ロボット学会誌, Vol.21, No.5, pp.480-483, 2003.

[11] 玉圭樹, 中谷一郎, ”深宇宙探査機の自律化とその検証”, 日本ロボット学会誌, Vol.21, No.5, pp.488-493, 2003.

[12] 浦環, ”水中に求められるロボット”, 日本ロボット学会誌, Vol.22, No.6, pp.692-696, 2004.

[13] 浦環, ”自律型海中ロボット r2D4 の制作と佐渡沖および黒島海丘海底観測”, 日本ロボット学会誌, Vol.22, No.6, pp.709-713, 2004.

[14] 近藤逸人, ”知的観測を行う水中ロボット”, 日本ロボット学会誌, Vol.22, No.6, pp.714-717, 2004.

[15] 柳善鉄, ”自律型水中群ロボットシステム”, 日本ロボット学会誌, Vol.22, No.6, pp.718-722, 2004.

[16] 金岡克弥, 川村貞夫, ”ダイバーロボットの実現に向けて”, 日本ロボット学会誌, Vol.22, No.6, pp.732-737, 2004.

[17] 安居院猛, 長橋宏, 高橋裕樹, "ニューラルプログラム", 昭晃堂, 1993.

[18] 銅谷賢治, "計算神経科学への招待〜脳の学習機構の理解を目指して〜", サイエンス社, 2007.

[19] T. Kohonen, Self-Organizing Maps, Spriner-Vcrlag, New York, 2001 third edition.

[20] Ian D. Kelly, David A. Keating, "Increased Learning Rates Through the Sharing of Experiences", Proceedings of the Seventh IEEE International Conference on Fuzzy Systems, 1998

[21] Ming Tan, "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents", In Proceedings of the Tenth International Conference on Machine Learning , 1993

[22] Ian D. Kelly, David A. Keating, Kevin Warwick, "Mutual Learning By Autonomous Mobile", Proceedings of the First Workshop on Teleoperation and Robotics, Applications in Science and Arts, 1997

[23] 木島康隆, "群の中の個体の知能の発達", 室蘭工業大学卒業研究論文, 2007.

[24] Yasutaka Kishima, Kentarou Kurashige, "Growth of individual intelligence using communication", Proceedings of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems, pp.287-292, 2008.

[25] Richard S. Sutton, Andrew G. Barto, "Reinforcement Learning", The MIT Press, 1998.

[26] Leslie Park Kaelbling, Michael L. Littman, Andrew W. Moore, "Reinforcement Learning A Survey", Journal of Artificial Intelligence Research 4, pp.237-285, 1996.

[27] 木村元, 宮崎和光, 小林重信, "強化学習システムの設計指針", 計測と情報, Vol.38, No.10, pp.618-623, 1996.

[28] J. Morimoto, K. Doya, "Acquisition of stand-up behavior by a real robot using hierarchical reinforcement learning ", Robotics and Autonomous Systems, Volume 36, pp. 37-51, 2001

[29] H. Kimura, T. Yamashita and S. Kobayashi, "Reinforcement Learning of Walking Behavior for a Four-Legged Robot", 40th IEEE Conf. on Decision and Control, pp.411-416, 2001

[30] M. Asada, E. Uchibe, and K. Hosoda, "Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development", Artificial Intelligence, Vol.110, pp.275-292, 1999

[31] Maja J. Matari, "Reinforcement Learning in the Multi-Robot Domain", Autonomous Robots, Vol.4, Number 1, 1997

[32] William D. Smart, Leslie Pack Kaelbling, "Effective reinforcement learning for mobile robots", Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference, 3404-3410 vol.4, 2002

[33] 伊庭幸人, ”統計学者・数理工学者のための統計物理入門　-格子スピン模型とマルコフ連鎖モンテカルロ法を中心として-”, 統計数理研究所, 1997.

[34] 伊庭幸人, 種村正美, 大森裕浩, 和合肇, 佐藤整尚, 高橋昭彦, ”計算統計 II マルコフ連鎖モンテカルロ法とその周辺（統計科学のフロンティア 12）”, 岩波書店, 2005.

[35] 浅田稔, 北野宏明, ”ロボカップ戦略: 研究プロジェクトとしての意義と価値”, 日本ロボット学会誌, Vol.18, No.8, pp.1081-1084, 2000.

[36] 田所諭, ”ロボカップレスキューリーグ”, 日本ロボット学会誌, Vol.27, No.9, pp.983-986, 2009.

謝辞

本論文を結にあたり，日ごろより懇切なるご指導を賜りました倉重健太郎先生に深く感謝の意を表します．また，ご助言，ご指導をいただいた畑中雅彦先生，本田泰先生，渡部修先生，渡邉真也先生に感謝の意を表します．そして，論文の査読や助言をしていただいた認知ロボティクス研究室の池田憲弘君，中南義典君，宮崎愛央君に感謝致します．

研究業績

[1] Yasutaka Kishima, Kentarou Kurashige, "Growth of individual intelligence using communication", Proceedings of Joint 4th International Conference on Soft Computing and Intelligent Systems and 9th International Symposium on advanced Intelligent Systems, pp.287-292, 2008.