知能と複雑系　121－14
（2000．7．19）

強化信号のコミュニケーションに基づくマルチエージェント強化学習

山口　智浩

奈良工業高等専門学校　情報工学科

〒639-1080　奈良県　大和郡山市　矢田町 22
TEL：0743-55-6140,　FAX：0743-55-6149,
E-mail：yamaguch@info.nara-k.ac.jp

あらまし　強化学習法は，様々な状況に柔軟に適応するエージェントの有力な学習メカニズムとして注目されている．しかしながら，エージェント間での相互依存性のあるマルチエージェント環境では，各エージェントへの適切な目標設定を行うのが困難であるので，他のエージェントらとの相互作用をどう調整するかを扱う，能動的かつ相互作用的な学習機能が必要となる．そこで本論文では，まず，強化学習エージェントの学習目標がコミュニケート可能な強化信号であるとみなし，エージェント間でそれらを相互作用的に生成，調整するための新しいマルチエージェント強化学習の枠組みを提案する．そして，学習の目標と評価基準の自律生成の課題について議論する．

キーワード　強化学習　マルチエージェント　自己反映　強化信号　コミュニケーション　インタラクティブ

Reinforcement Signal Communication based Multiagent Reinforcement Learning

Tomohiro YAMAGUCHI

Department of Information Engineering, Nara National College of Technology
22, Yata-cho, Yamato-Koriyama-city, Nara, 639-1080, Japan
Tel: +81 -743-55-6140,　Fax: +81 -743-55-6149,
E-mail: yamaguch@info.nara-k.ac.jp

Abstract　Reinforcement learning is the major learning mechanism for an agent to adapt itself to various situations flexibly. However, in a multiagent system environment that has mutual dependency among agents, it is difficult for a human to setup suitable learning goals for each agent. Therefore, it requires the active and interactive learning function that treats how to coordinate the interaction among other learning agents. This paper presents a new framework of multiagent reinforcement learning to generate and coordinate each learning goal interactively among agents. To realize this, it presents to treat each learning goal as a reinforcement signal that can be communicated among agents. Then the issues of the self-generation of goals and evaluation criteria are discussed.

key words　reinforcement learning, multiagent, self-reflection, reinforcement signal, communication, interactive

1. はじめに

AI の主たる目標の一つに，自己の価値観を持って自律的に行動する知的エージェントの実現がある．AI 研究のスタイルが，組み込み知識を持つ閉じた AI システムから，外界と相互作用しあうエージェント[3]へと移りつつある．ここでエージェントとは，複雑，動的な環境下で与えられた目標の達成を試みるシステム [9] である．そのため，変化する外界へ適切に対処するための適応的で柔軟なエージェントの自律性をどう実現するかが論点[9]となっている．

動的な状況下では，あらかじめ状況の特定，想定が困難なため，従来型 AI のように，必要な知識を設計者がシステムやエージェントに用意することができない．そのため，自律的エージェントには，直面した外界の状況に応じて目標達成に必要なスキルを実行時に獲得しながら行動する，経験からの学習機能が要求される．

強化学習[1, 3, 14, 15, 16]は，未知の環境に対する自律エージェントの学習機能として有望な手法のひとつであるが．学習の適応的で柔軟な自律性については，未解決の問題がある．

まず．強化学習での学習目標，および評価基準は，通常設計者である人間から与えられるため，これまでの大半の研究では，それらをエージェントが自身の行動する環境に対し，自律的にどう決定するか，という問題は議論されてこなかった．しかしながら，学習の視点を設計者からエージェントへ移して考えると，これらは，強化学習効率を左右するだけでなく，環境の変動にどう能動的に適応するか，という重要な課題である．

さらに**マルチエージェントシステム**（MAS と略す）[2, 19, 20, 21]は，学習エージェントにとって典型的な複雑，動的な環境である．そのため設計者が MAS 全体に与えた目標に対し，各エージェントがどう適切に各自の目標設定を行うかは，両者にとって困難な問題である．MAS における，他エージェントとのインタラクティブな環境下で，自律的エージェントを実現するためには，以下の相互に関連した３つの課題（副問題）を解決する必要がある．

（１）目標，評価基準の自律生成 [13]
（２）エージェント自身の価値観の形成 [10]
（３）自他の領域・境界の設定 [11]

これらのうち，本論文は，主に（１）を対象とする．強化学習の枠組みをベースとした自律的エージェントにおいて，まず，強化学習法を MAS に拡張するための手法について提案し，学習における目標と評価基準の自律生成の課題と解決手法について議論する．ここで評価基準とは，複数の解や方法がある場合に，それらの間で最適なものを決定，順序づけるための知識である．

では，本論文の構成を以下に示す．２章では，強化学習法の概要と強化学習を MAS に適用する場合の課題について述べ，エージェント間での相互依存性のある MAS の場合，集団での各エージェントの自己利益追求の和は集団での利益最大化にならない，という社会的ジレンマについて議論し，これを解決するための**マルチエージェント**（MA と略す）間での学習の相互作用的な調整機能の必要性を明らかにする．３章では，まず強化学習における学習目標を，MA 間でコミュニケート可能な強化信号へと拡張することについて議論する．次に学習の方向付けの自律性について，以下の２つの課題があり，それらが矛盾することを示す．

１）学習の客観性，合理性を保証するには，学習目標はエージェントの外部かつ固定であるべき．
２）外的な報酬・目標は，学習者の内発的動機付けを低下させる．

これを解決するため，強化信号に基づく**マルチエージェント強化学習**（MARL と略す）の枠組みを提案する．そこでは，各エージェントが自他の学習目標を強化信号として生成，発信し，エージェント間で強化信号を通信しながら

１）自己の内発的な目標設定
２）他者からの外的な強化刺激の利用，

との両者のバランスを相互に調整することによって上述の矛盾の解消を図る．

４章では学習目標と評価基準の自律生成の課題について議論する．まず，大局的外部目標が存在する場合には，それと矛盾せず，社会的ジレンマを減少させる制約が，各エージェントの自律的目標生成の合理的な評価基準であることを示す．次に，大局的外部目標が存在しない場合には，強化信号に基づく MARL において，各エージェントの個別に探索，生成する学習目標の生成，評価基準が，MAS 内で共有されることが，各エージェントの獲得報酬和を最大化し，かつ共有された評価基準が進化的に安定であることを示す．

2. 理論

本章では強化学習の基本的な枠組みと代表的手法の特徴について述べ，従来手法をマルチエージェント強化学習に適用する場合の問題点について議論する．

2.1 強化学習法の概要

本節では，まず強化学習の基本的な枠組みについて述べ，次に代表的な２つの強化学習手法の違いについて議論する．

2.1.1 強化学習の基本的な枠組み

強化学習は，外界（環境と呼ぶ）との入出力を通して相互作用しながら行動する**エージェント**（agent，**主体**と呼ぶ）を基本的な枠組みに持つ．主体は，環境からの感覚入力（**状態**と呼ぶ）に対して**行為**（action）を選択，実行する．行為の結果は主体の状態の変化として環境から返され，**報酬**が設定された状態に主体が遷移すれば報酬が与えられる．強化学習法の利点は，目標状態を報酬で指示するだけで，環境に応じて任意の状態から目標状態に至る最適な行動系列が学習で得られる点である．

環境を確率的状態遷移グラフとしてモデル化した場合，各状態からの遷移を区別するラベルが行為である．状態に対する行為のペアを**行動**（単に**ルール**と呼ぶ），全ての状態に対し選択すべき行動の集合を**政策**（policy）と呼ぶ．学習の目的，すなわち学習結果の評価基準は単位行動当たりでの期待獲得報酬の最大化で，これを最大化する行動を各状態で与える政策を**最適政策**と呼ぶ．期待獲得報酬は政策に依存するので，主体は，環境と相互作用しながら政策を探索することになる．この環境との相互作用をどう解釈するかで，強化学習研究は客観観測型と主観経験型との２種に分類できる．

2.1.2 客観観測型学習 vs 主観経験型学習

客観観測型学習の**観測**とは，主体を環境から切り離した上で環境をモデル化することで，その特徴は，政策に依存しない観測による環境モデル同定と，モデル上での最適政策の網羅的探索とに分けて学習する点である．環境のモデル化を行うには，環境のクラスを仮定する必要がある．意志決定理論での動的計画法（DP）では，マルコフ決定過程（MDP）モデルでの最適化手法が知られていることから，観測結果からMDPモデルを統計的に同定し，最適政策をDP法で探索するad-DP法[1]が有名である．客観観測型は，環境のクラスがMDPやその類似クラスに限定される反面，最適性を追求できる利点[16]がある．

これに対し主観経験型学習の**経験**とは，行動や政策に依存した学習結果を指す．主観経験型には，バケツリレー，Profit Sharing，Q 学習法などモデルレスの強化学習法が相当する．手法によるが，学習結果が得られた経験の順序や分布に影響を受けるため，通常は最適解に収束しない．例えばQ学習法が有限回で最適政策に収束しないのは，強化に用いる経験が実行した政策に依存し，かつ異なる政策を混合して学習するからである．むしろ主観経験型の主張は，主体の行動による相互作用は環境を変化させるので，客観観測，環境のモデル化は困難，という点である．したがって最適性を追求しない手法が多い．

2.2 MDPベースの強化学習手法の問題点

本節では，まず客観観測型強化学習の代表的手法である，MDPベースの強化学習の特徴とそれをMASに適用する場合の問題点について議論し，次にその解決策である主観経験型強化学習について述べる．

客観観測型強化学習は，MDPベースの強化学習手法で実現される．その特徴は，観測の客観性と完全性の仮定である．前者の観測の客観性とは，観測による行動は環境を変化させず，環境との相互作用なしと仮定する点である．後者の完全観測とは，環境中の全ての状態は，区別／観測できるという性質である．これによって主体は，環境を完全にモデル化可能である．つまり，MDPベースの強化学習手法の特徴は，主体を環境から切り離した上で，観測者の視点からの客観的観測によって環境を完全にモデル化しようとする点である．

しかしながら，MDPベースの強化学習手法の問題点は，マルチエージェント強化学習への適用が原理的に困難なことである．1990年代後半において，非マルコフ環境であるMASでの強化学習研究[2,20,21]が盛んになりつつある．MAS環境が重要なのは，マルチエージェント間の相互作用によって生じる環境の動的性と客観観測型学習とが両立しない世界だからである．すなわちMASでは，主体同士が相互作用するので，学習主体が客観観測する限りMASの構成要素とはなれず，一方MASの一員となって行動し他の主体と相互作用すれば，それがMAS環境に影響を与えるため，観測が主観的経験となるからである．さらに学習主体同士では相互のモデル化の無限退行という問題も発生する．したがってMAS環境では，主観経験型同士の相互作用を扱う相互主観的な強化学習研究が望まれる．

そこで，行動主体の視点からの学習である主観経験型強化学習について議論する．その特徴は，観測の主観性と部分性の仮定である．全者の観測の主観性とは，観測と行為とが独立でないため，主体の行為に観測結果が影響を受けてしまう性質である．後者の観測の部分性とは，エージェントの視点では，他のエージェントの内部状態などの見えない部分が存在する性質である．本論文では，**経験**とは，主体の視点からの，主体の行動に依存した環境との相互作用の主観的な観測結果，と定義する．

以上より，MAS環境では，既存のMDPベースの強化学習手法が仮定する客観的，完全観測は困難であるため，主観的かつ部分観測による経験からの強化学習手法が必要となる．

2.3 マルチエージェント強化学習の課題

本節では，まずマルチエージェントの強化学習（MARL）研究[2]の困難な点について述べ，次にMARL

の論点と問題点について説明する．
まずマルチエージェントの強化学習研究の困難な点は，

1）複数の学習主体が存在することによって生じる環境の非決定性
2）主体の行動選択に依存した主観的経験からの学習の問題

である．MAS では非集中制御を基にするため，1）と2）とを主体視点では独立に扱えない．したがって各学習エージェント視点では，非マルコフ環境となる．つまり，主体間の相互依存性があるため，学習の最適性の定義が困難なことである．一方，1）2）とを客観的視点で扱う分野は，既存のゲーム理論であるので，客観観測主義強化学習では，高々ゲーム理論と同程度の知見しか得られない．
次にMASのゲーム理論的解析での問題点[19,20]について述べる．第1に主体間での利益分配問題[18]について述べる．これは，集団での目標達成で得た報酬を各学習主体にどう分配すべきかという，エージェント間の信頼度割り当て問題（CAP）と呼ばれる．その論点は，合理的な分配方法は存在せず，何が公正かという価値観の問題であることである．つまり，CAPのための客観的な合理性の定義が困難なので，なんらかの形での主体間での利益分配や利害の調整，価値観の形成機能が必要[18]である．
第2は，各個体での自己利益の和が，集団での利益と一致しない問題である．なぜなら，各エージェントにとって学習の基本原理は，報酬獲得という形での自己利益最大化である．しかしながらエージェント間での相互依存性のあるMASの場合，集団での各エージェントの自己利益追求の和は集団での利益最大化にならない，という**社会的ジレンマ**[19,20]が生じるからである．言い換えると，各強化学習エージェントの最適化学習である個人の自己利益追求が，MAS全体の損失となり得ることでもある．これは，集団挙動において個人の行動の貢献度をどう適切に評価するか，という信頼度割り当て問題に原因がある．つまりMARLにおいては，各個体の効用最大化と集団での効用最大化とが矛盾しないように学習目標である報酬を設定する必要があるが，社会的ジレンマを生じないような事前の報酬設定が，一般には容易ではない．なぜならエージェント間での相互依存性のあるMAS環境では，各エージェントへの適切な目標設定を行うのが困難だからである．しかも相互依存性を解消せずに最適政策を全探索で求める従来の客観観測型強化学習手法をそのままMAS環境に適用するのは，計算量が膨大となって非現実的である．そこで，他のエージェントらとの相互作用をどう調整するかを扱う，能動的かつ相互作用的な学習機能が必要となる．これを近似的に解決する有力な方法は，各エージェントの自己利益最大化の和が集団での利益最大化となるように，つまり社会的ジレンマが小さくなるように，エージェントレベルでの目標設定を動的に調整することで，エージェントごとに独立な部分問題に分割し，個別に強化学習することである．
これに関連して利他的行動の非合理性の問題がある．これは，社会的ジレンマを解決する利他的行動の合理性は何か？ということである．では，これらの問題を解決するためのマルチエージェント強化学習の論点を要約する．まず，集団タスクに設定した報酬をエージェント間に分配する信頼度割り当て問題がある．次に，エージェント間の相互依存性を解消する方法の一つとして相手モデルの学習がある．第3に，エージェント間での利害の調整，意味の共有，役割分化，大域的な社会性ルールの発生などのためにMA間でのコミュニケーションの発達が重要である．エージェント間の信頼度割り当て問題を解決するため，次章では，強化学習エージェントの学習目標をエージェント間で相互作用的に生成，調整するための新しい強化学習の枠組みを提案する．

3. 相互作用的自己反映に基づく強化学習法

本章では，まず強化学習法における強化信号の役割について再検討した上で，強化信号のコミュニケーションに基づくマルチエージェント強化学習の新しい枠組みを提案する．

3.1 強化信号と報酬との起源と役割

強化学習とは，強化信号という学習のフィードバック情報を用いる機械学習法の一種である．**強化**（reinforcement）とは，元々は，動物の行動のメカニズムを刺激に対する反応で理解しようとした1910年代の行動主義心理学において，動物が，エサ等の特定の刺激（**強化刺激**と呼ぶ）に対して，次第に特定の反応を繰り返すようになることを指す．これに対しAIでの強化学習とは，行動型（behavior-based）AIシステムにおいて，強化信号を伴う入力を数多く得るような行動出力を獲得するための入出力の写像を学習するメカニズムを指す．一般に**強化信号**はスカラー量で表され，正の信号を**報酬**（reward），負の信号を**罰**と呼び，罰を回避し報酬を得る行動を強化するように学習が進む．したがって目標状態に報酬を設定すれば，目標状態に到達する最適な方法を強化学習によって得ることができる．
行動科学分野では，学習者の行動を左右する強化刺激とは環境中の何か？が追求されてきたのに対し，これまでの強化学習研究では，OR分野での効用の扱いに関する影響もあって報酬とは何か，といった基

本的問題に関する議論が少なかった．しかしながら，報酬から強化信号へと発想を転換すると，環境中に存在する信号の一部を学習を方向付ける強化信号として利用することが可能となる．

3.2 なぜ強化信号の自律生成なのか？

本節では，学習の方向付けの自律性についての２つの課題について議論する．

第１の課題は，強化学習において外部報酬を正当化してきた学習基準の自己評価の問題である．従来の強化学習法の枠組みにおける報酬の役割と問題点について議論する．強化学習エージェントにとって報酬とは何か？報酬とは，エージェントの外部にあって，エージェントの学習を導く固定目標であり，設計者が与えるものである．なぜなら，機械学習の分野では，学習基準の自己評価の問題があるためである．つまり学習目標や学習基準がエージェント内にあって操作可能である場合には，エージェントの得た学習結果を正当化するように，学習目標を変更するおそれがある．そのため，学習目標は，固定かつエージェントの外部にある必要があると考えられてきた[3]．

しかしながら心理学分野では，外部報酬は人間のやる気を高めると考えられてきた常識に対し，1970 年代に人間において外的報酬は学習者の内発的動機付けを低下させる，という新たな知見が提案[4-6]され，実験的に裏付けられている[7]．ここで，内発的動機付けとは，人間におけるやる気，意欲などを表し，学習者自身が発見，形成していく自律的な学習目標のことである．第２の課題は，強化学習における自己の学習目標生成の問題である．心理学の知見を機械学習に当てはめて考えると，既存の外部からの固定報酬による強化学習手法は，エージェントにおける学習の自律性に対しては負の要因である．つまり，学習の方向性は，外部の報酬設計者に委ねられたままで，エージェント研究における自律性の実現にとっては，むしろ障害となっている．

さらに，既存の強化学習研究では，適切な報酬の設計法に関する理論が明らかでないため，単純な学習問題では，報酬設定が容易なので学習効率がよいが，複雑な問題では，適切な報酬設定である学習の方向付けを設計者もエージェント自身も行えないため，学習効率が悪化する，という本質的な問題を抱えている．したがって，２章で議論した MA 学習における論点を解決するためには，まず人間の内発的動機付けに対応する，学習エージェント自身による自律的な学習目標設定のメカニズムの実現が重要である．

さらに，学習基準の自己評価の問題を回避するアイデアとして，MA 間での相互評価による学習目標の修正メカニズムを導入する．そこで両者を統合した新しい MA 強化学習の枠組みとして，次節では，強化信号の自律的選択・生成・コミュニケーションによる強化学習法を提案する．

3.3 強化信号のコミュニケーション

本節では強化信号のコミュニケーションに基づく強化学習の新しい枠組みを３つのステップに分けて説明する．

3.3.1 強化信号の自己生成による，学習目標の能動的生成

学習主体が，内部に強化信号を生成する**報酬生成関数**（Payoff 関数）を持ち，自己および他者の行為に対する強化基準として用いる．各状態でのエージェントの実行可能な全行動ルールに対して，報酬を与えるかどうかを記述する報酬生成関数を本研究ではテーブル表現で表し，これを**報酬テーブル**と呼ぶ．学習エージェントは，報酬テーブル上の各行動ルールに対して，報酬の大きさを設定，修正する機能によって，自己の学習目標を探索すると共に，自己が実行したルールに対して，報酬テーブルで設定された報酬を用いて強化学習を行う．

3.3.2 強化信号の自律的選択による能動的強化学習

学習主体が強化基準を内部に報酬テーブルの形で持つと，次に問題となるのが学習目標を記述，誘導し，方向付ける報酬をどう設定するかである．MA での集団目標のように人間によって与えられた固定的な外部報酬がある場合には，それと矛盾しない副報酬を自己生成すればよい．それ以外の場合には，学習主体が，環境中から強化信号を選ぶ基準を学習主体に与えることによって，学習目標を主体自らが決め，報酬テーブル上に設定するようにする．

用いる強化信号の候補は，他者が生成した強化信号または外部報酬と共起，関連する環境中の情報が利用可能である．本論文では，自己が外部報酬を得た，あるいは得られなかった経験を手がかりに強化信号を生成する．

3.3.3 強化信号の送受信による，学習目標の相互作用的自己反映

エージェントがなんらかの行為を実行した時に、そのエージェントを含めた環境中の全エージェントは、その行為に対する報酬を自己の報酬テーブルから生成し、強化信号として行為を実行したエージェントに送信する．このとき、行為を実行したエージェント自身が生成する強化信号を**自己強化信号**，それ以外の他エージェントから送信された強化信号を**他者強化信号**と呼び、それぞれを区別する．行為を実行した結果，他者から強化信号を受けたエージェントが，他者強化と自己強化の両方を統合して報酬 Rw(i)

を生成する式（1）を示す.

$$Rw(i) = G\_Rw + \alpha * S\_RS + \beta * \sum E\_RS \quad \cdots \quad (1)$$

G_Rw：（人間が設定した）外部固定報酬
S_RS：自己強化信号、E_RS：他者強化信号
α：自己強化の信頼度、β：他者強化の信頼度
（$-1 \leqq \alpha, \beta \leqq 1$）

式（1）中のRw(i)を用いて，自己の学習テーブルを更新した後に，自己の強化基準である学習テーブル更新式中のパラメータα、βで内部強化と外部強化のバランスを修正することで、相互依存問題の解消を目指す.
本研究の着眼点は，強化学習における学習目標を表す報酬を，特別な情報としてではなく，環境にある情報の一部としての強化信号へと発想を変えた点である．強化信号とは，環境に対して，エージェントが知覚，出力する情報の一部である．各エージェントが，学習に用いる強化信号を感覚入力情報から自律的に選択することで，自己の学習の目標づけが可能となる．さらに，自己の内部状態の一部を，環境に（強化）信号として出力することで，他者の行動学習の制約／誘導が可能となる.

4. 学習における目標と評価基準の自律生成

4章では学習目標と評価基準の自律生成の課題について議論し，強化信号に基づくMARLにおいて，各エージェントが個別に探索，生成する学習目標の評価基準が，MAS内で共有されるためのシナリオを示す.

4.1 強化学習における評価基準の分類

一般に学習の評価基準は，評価する対象によって，以下の3種に分類できる．これらは全て通常，設計者である人間によって設定される.

1．**学習目標**の評価基準：学習目標は，エージェントが獲得すべき学習結果を表す制約である．目標は，問題に対し人間にとって自明である場合が多く，それゆえ，与えられた問題空間に対し．学習目標をどう選ぶかを決める評価基準を明示するのは困難である.

2．**学習過程**の評価基準：学習過程とは学習空間において，解を表す学習結果を求める探索のことで，これを制御する知識は，強化学習ではエージェントの行動選択戦略と呼ばれ，これによってエージェントが環境を探査する順序が決まる．これには，大きく分けて学習過程での獲得報酬和を重視するExploitation手法と，環境の同定を重視するExploration手法との2つがある.

3．**学習結果**の評価基準：これは，求める学習結果の質を決める基準である．強化学習の場合，学習結果は，学習目標を達成するための政策である．通常は，政策の評価基準は，期待割引報酬和の最大化が用いられる.

これら3つの評価基準には，依存関係があり，4.4節で詳しく議論する.

4.2 学習目標の評価基準としてのエージェントの自己コントロールルール

本節では，各エージェントが生成する自他に対する強化信号の生成規則，すなわち学習目標の評価基準を，心理学の用語を用いて**自己コントロールルール**[8, 14]と呼ぶ.
通常の強化学習では，学習目標は報酬で表され，目標状態には報酬を設定し，エージェントが目標状態に遷移すると報酬が与えられる．これに対し，強化信号に基づくMARLでは，学習目標は，自他強化学習エージェントが，報酬テーブルを用いて生成，発信する強化信号の重み和で表される．ここで問題となるのが，

1）報酬テーブル上で，どの行動ルールに強化信号を設定すべきか，および，
2）学習に用いる強化信号の重みの調整の決め方，

という2種類の強化信号生成の評価基準の選び方である．上述の**自己コントロールルール**は，これらを合わせたものである．そして，自己コントロールルールを探索，修正するしくみをエージェントに与えることにする．つまり，自己コントロールルールとは，自己および他者の行動の強化の仕方を決めるメタ規則である.

4.3 外部報酬下における学習の目標と評価基準の自律的生成

4.3節では，学習結果と学習目標生成の評価基準との同時学習の問題について議論する．設計者が設定したMAS全体に対する大局的外部目標が存在する場合には，これと矛盾せず，社会的ジレンマを減少させる強化の方向が，各エージェントの自律的目標生成に対する設計者の視点での合理的な評価基準であることを提案する．特に，3章で提案した強化信号に基づくMARLの場合，社会的ジレンマを解決する解，すなわち，MAS全体の報酬が，各エージェントの報酬和と一致する，という制約が各エージェントの自己コントロールルール（強化信号の自律生成および，自他の強化信号の重みの調整）の収束条件を表す合

理的な評価基準であることを述べる．
まず，大局的外部報酬の獲得和の最大化を学習結果の評価基準として与える．次に，各エージェントへの学習目標生成の評価に用いる必要条件的な指標として，社会的ジレンマ度を以下のように定義する．

社会的ジレンマ度 ＝（MAS 全体で獲得する大局的外部報酬の和）－（各エージェントが獲得した自他強化信号の全エージェントでの和）

その時，各エージェントの学習目標を表す強化信号生成の評価基準は，社会的ジレンマ度の最小化，が合理的である．すなわち各エージェントは，社会的ジレンマ度を減少させる方向へ各エージェントの学習目標，すなわち強化信号生成のための自己コントロールルールを修正する．そして，以下の２つの項目について，

1）社会的ジレンマ度が減少し，生成強化信号の質が改善される．
2）MAS での大局的報酬の獲得和が増加する．

1）と2）とが相互に改善を促進しあう正のフィードバックが生じると，学習結果と学習の副目標の生成との同時学習の問題が解決される．

4.4 外部制約がない場合での学習の合理性

本節では，学習の合理性，外部制約，効率化の３つの関係について議論する．
前節では，外部目標がある場合での学習目標と評価基準の自律生成の課題について議論した．では，この議論の背景にある，基本原理について述べる．
4.1 節で検討した学習の評価基準の合理性は，外部制約の下での行動の効率化追求という基本原理で定義，説明できる．エージェントが行動学習する場合に，制約となるのは，行動に要する何らかの資源（時間コスト，空間コスト，etc.）についての制約や競合である．この場合，その制約の下での資源利用の効率最大化（例えば，学習コスト最小）が学習の合理性定義の基本原理である．
したがって，学習目標の自律生成の場合，外部制約を表す外部目標が与えられない場合，目標の自律生成の合理性は，客観的には定義できない．したがって，エージェントが目標を自律生成するための評価基準は，主観的である．そこで外部制約が存在しない場合には，MAS で共有する外部制約を形成できれば，その制約の下での各エージェントの学習の合理性が定義可能となる．つまり，MAS での学習の合理性は，MAS で共有される相互主観的な制約で定義される．そこで，次節では，外部制約が存在しない場合での MAS における，自己コントロールルールの相互主観的な自律形成について議論する．

4.5 社会的強化の発達

本節では，大局的目標が存在しない場合に，自律的な目標や評価基準を表す自己コントロールルールが，MAS 全体で形成されるシナリオについて議論する．
まず，MAS に対し，大局的な外部報酬が存在せず，各エージェントは，エージェントが生成する自他強化信号の和のみによって，強化学習を行うとする．このような状況に対し，社会学では，現在の人間社会では，自己・他者の行動の規範となる自己コントロールルールが社会に広まり，共有されるようになると，社会で広まっている自己コントロールルールへの同化が同化しない場合より有利なため，社会での自己コントロールの共有が促進されるとともに自己コントロールがさらに強化される[14]，と指摘している．
これを MARL に適用して考えると，ある自己コントロールルールが進化的安定状態になる条件は，以下の２つの要因について，正のフィードバックが生じることである．

1）自己コントロールルールの共有
2）自己および他者からの強化信号獲得和の最大化

ここでの自己コントロールルールとは，例えば，あるエージェントが獲得した大局的報酬を同一の自己コントロールルールを持つエージェントのみに分配する，といった互恵主義的ルールを想定[12]している．エージェントの学習結果の評価基準は，自己の利益の最大化であるので，それを表すエージェント当たりの報酬獲得和を比較したときに，ある自己コントロールルールを共有するエージェント群が，他の自己コントロールルールを持つエージェント群よりも，エージェント当たりの報酬獲得和が大きい場合には，前者の自己コントロールルールが進化的に有利かつ安定になるので，それが次第に MAS 全体に広がっていく．
自己コントロールルールが MA 間で共有され，エージェント間相互で強化信号をやりとりしながら，強化学習を行う状況を**社会的強化信号の発達**と呼ぶ．社会的強化信号が発達した MAS では，学習目標の評価基準の合理性は，以下の２つの要因で表されている．

1）自己コントロールルール自体の合理性（各エージェントの獲得報酬和）
2）自己コントロールルールの共有の度合い

すなわち，MAS で収束する自己コントロールルールの質は，各エージェントの獲得報酬和で評価され，かつ自己コントロールルールの共有度が高まるほど，各エージェントの獲得報酬和が改善される，という

のが，本論文で提案する社会的強化信号が発達するシナリオである．

5. 結論

本論文では，強化学習エージェントの学習目標をエージェント間で相互作用的に生成，調整するための，MA における強化学習の新しい枠組みを提案した．次に，学習の目標と評価基準の自律生成の課題について議論した．そして，MAS における自己コントロールルールの共有が，各エージェントに対し学習目標の自律生成のための適切な評価基準を導くことを述べ，それが MAS で共有され，社会的強化（規範）となるためのシナリオについて述べた．本手法の利点は，報酬の初期設定が最適でない場合に，各学習エージェント間で利害調整を行うことによって，報酬設定を改善しながら強化学習できる点である．
本手法は，MA のパフォーマンス改善に広く応用可能である．今後の課題は，絶対的な外部報酬が存在しない場合の適切な学習基準を実装することである．

References

[1] Barto, A.G., Bradtke, S.J. and Singh, S.P., Learning to act using real-time dynamic programming, Artificial Intelligence 73, pp. 81-138, 1995

[2] Sen, S. and Weiss, G., Learning in Multiagent Systems, in Multiagent Systems - A Modern Approach to Distributed Artificial Intelligence, G. Weiss (Ed.), pp. 259-298, The MIT Press, 1999

[3] Russell, S. and Norvig, P., Artificial Intelligence - a modern approach -, pp. 525-529, Prentice-Hall International, Inc., New Jersey, 1995

[4] Deci, E.L. and Flaste, R., Why We do What We do - The dynamics of personal autonomy -, G.P. Putnam's Sons, 1995

[5] Deci, E.L., Effects of externally mediated rewards on intrinsic motivation, J. of Personality and Social Psychology, 18, pp. 105-115, 1971

[6] Deci, E.L., Intrinsic motivation extrinsic reinforcement, and inequity, J. of Personality and Social Psychology, 22, pp. 113-120, 1972

[7] Lepper, M., Greene, D. and Nisbett, R., Understanding children's intrinsic interest with extrinsic reward: A test of "overjustification" hypothesis, J. of Personality and Social Psychology, 28, pp. 129-137, 1973

[8] Goldman, D., Emotional Intelligence, Bantam Books, 1996

[9] Maes, P., Modeling Adaptive Autonomous Agents, J. of Artificial Life, Vol. 1, No. 1/2, pp. 135-162, 1994

[10] Shiose, T., Sawaragi, T., Katai, O. and Okada, M., Autonomy from the viewpoint of Bi-Referential Model, S The Third Asia-Pacific Conference on Simulated Evolution And Learning (SEAL2000), Oct. 2000 (to appear)

[11] Takadama, K., Inoue, H. and Shimohara, K., How to Autonomously Decide Boundary Between an Agent and Others?, The Third Asia-Pacific Conference on Simulated Evolution And Learning (SEAL2000), Oct. 2000 (to appear)

[12] Yamaguchi, T., Kitahashi, M. and Yachida, M., The Species Fitness Method for the Evolution of Cooperative Behavior in a Group Task, Journal of Artificial Life and Robotics, Springer-Verlag, vol. 3, pp. 127-132, 1999

[13] Yamaguchi, T. and Watanabe, R., Interactive Self-Reflection based Multiagent Reinforcement Learning for Coordination, The Third Asia-Pacific Conference on Simulated Evolution And Learning (SEAL2000), Oct. 2000 (to appear)

[14] Sutton, R.S., and Barto, A.G., Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998

[15] Kaelbling, L.P., Littman, M.L., and Moore, A.W., Reinforcement Learning: A Survey, Journal of Artificial Intelligence Research, Vol. 4, pp. 237-277, 1996

[16] 山口 智浩，石村 健二，RAE－PIA：報酬獲得効率を最大化する政策の強化学習，第 14 回人工知能学会全国大会論文集，July, 2000

[17] 森 真一，自己コントロールの檻，-感情マネジメント社会の現実-，講談社選書メチエ，2000

[18] 竹内 靖雄，経済倫理学のすすめ，-感情から勘定へ-，中公新書，1989

[19] 生天目 章，マルチエージェントと複雑系，森北出版，1998

[20] 伊藤 昭，ゲーム理論とマルチエージェントシステム，人工知能学会誌，Vol. 12, No. 2, pp. 223-230, 1997

[21] 三上 貞芳，強化学習のマルチエージェント系への応用，人工知能学会誌，Vol. 12, No. 6, pp. 845-849, 1997

盛光印刷所