	マツナミ　ナツキ
氏　　名	**松波　夏樹**
学位の種類	博士（工学）
学位記番号	博第1249号
学位授与の日付	2022年3月31日
学位授与の条件	学位規則第４条第１項該当　課程博士
学位論文題目	マルチエージェント強化学習におけるチーミング機構に関する研究 -連携を促進する組織構造と報酬メカニズムの設計- (A Study on Teaming Mechanisms in Multiagent Reinforcement Learning -Designing Organizational Structures and Reward Mechanisms that Promote Collaboration-)
論文審査委員	主査　教授　加藤　昇平 教授　白松　俊 准教授　松井　俊浩 准教授　大塚　孝信 教授　伊藤　孝行 (京都大学)

論文内容の要旨

強化学習は近年目覚ましい発展を見せているが，いまだ実用に向けては様々な課題がある．その代表的な例が学習するエージェントが環境中に複数存在するマルチエージェント環境の考慮である．

本研究の目的は，動的及び競争的環境への適用を念頭に，様々な環境複雑性を有する問題を対象に，マルチエージェント強化学習(Multiagent reinforcement learning；MARL) によって望ましいチームワークを実現するエージェントを得る方法を明らかにすることである．学習エージェントによるチームワーク実現のために，本研究では大きく分けて次の 2 点の提案を行う．

提案 1. チームを構成するエージェントの組織構造と学習方式に対する工夫

提案 2. 報酬メカニズムの設計

まず提案 1 について，MARL による解決を困難とする要因である連続空間，部分観測情報環境，競争的環境といった条件について整理し，これらの特徴を全て持つ過酷な環境での追跡問題を対象に問題設定を行う．そのうえで，環境困難性を緩和するチームの組織構造と，学習方式による対処方法として具体的には次の 2 点の提案を行う．1 つ目は，チームにおいて能力に優れたものが Leader となって他のエージェントに指示を行う Leader-Follower モデルの導入と Leader から Follower に対する一定の強制力を持った通信を付与することであり，2 つ目は，学習の初期段階において競争的環境にある一方のチームに対しもう一方のチームが「あえて負ける」行動を行うカリキュラム学習と，Train and

evaluation による学習フレームワークの工夫である．提案手法の有効性を確認するため，追跡問題に対する学習を複数の手法を用いて実験し，従来からある他の手法と比較して提案手法の有効性を確認するとともに，提案手法を構成する各要素毎の影響についても分析する．

提案 2 は，強化学習においてエージェントの方策を特徴づける報酬設計について議論する．複数のエージェントが同時に学習する MARL では，エージェントの自律性を損なうことなく分権的に，全体として好ましい協調を実現することが望まれる．複数のエージェントが行った共同行動の結果に対する報酬だけではなく，個々のエージェントの貢献度に応じた報酬信号を設計することができれば，学習エージェントは容易に全体にとって望ましい行動を行うような学習を実現することができる．しかし，協調タスクにおけるインセンティブとしての報酬設計はエージェントの貢献度合に応じて，成果の分配を決める貢献度分配問題(credit assignment problem)に帰着し，協調すべきエージェント達の誘因を損なうことなくシステムの要求目標を実現するような報酬関数を設計することは容易ではない．一方メカニズムデザインでは，ミクロ経済学とゲーム理論の一分野であり，複数の利己的なエージェントをいかにして効率よく取りまとめるかという問題を扱い，社会的余剰が最大となるような設計を行う．本研究では，メカニズムデザインの一例として Vickrey-Clarke-Groves メカニズムによる支払いのルールである迷惑料の考え方に基づいて，個々のエージェントが仮に存在しなかった場合の社会の効用の差分に基づいてそのエージェントに対する報酬を計算する手法を提案し，実験を行って評価した．

VCG メカニズムでは，評価対象のエージェントの貢献を評価するために，そのエージェントが存在しなかった場合の外部性を評価する必要がある．VCG メカニズムに基づく支払いによる報酬設計の適用可能性を拡大させるため，エージェントの不在性評価が容易ではない問題であっても適用可能な方法として，評価対象のエージェントが存在しない仮想環境を用いた MARL 手法についても提案する．2 種類の問題設定を対象に学習を行って結果を評価する実験を行い，各種従来手法と比較して議論する．

以上から，本研究では 1.チームを構成するエージェントの組織構造と学習方式に対する工夫 及び 2.報酬メカニズムの設計 という大きく 2 点の提案を行い，従来課題であった環境複雑性による学習困難性の緩和と，学習エージェント間のインセンティブ設計を反映した報酬設計による協調の発露について実例を示す．

本研究の成果は，人間がそう遠くない将来に直面する，学習によって駆動する多数のエージェントと共生する社会，すなわち AI エージェント同士，あるいは人エージェント同士，さらには人及び AI エージェントが混然一体となった社会状況において，互いに望ましいチームワークを実現するための中央集権性と分権性（Centralized/De-centralized）のあり方について，組織構造と報酬メカニズムの設計という側面からの新たな知見と，今後の可能性を示している．

論文審査結果の要旨

強化学習は近年目覚ましい発展を見せているが，いまだ実用に向けては様々な課題がある．その代表的な例がマルチエージェント環境の考慮である．本研究の目的は，動的及び競争的環境への適用を念頭に，様々な環境複雑性を有する問題を対象に，MARL によって望ましいチームワークを実現するエージェントを得る方法を明らかにすることである．学習エージェントによるチームワーク実現のために，本研究では大きく分けて次の 2 点が提案されている．

1. チームを構成するエージェントの組織構造と学習方式に対する工夫
2. 報酬メカニズムの設計

まず 1.について，MARL による解決を困難とする要因である連続空間，部分観測情報環境，競争的環境といった条件について整理し，これらの特徴を全て持つ過酷な環境での追跡問題を対象に問題設定を行う．そのうえで，環境困難性を緩和するチームの組織構造と，学習方式による対処方法として具体的には次の 2 点の提案を行う．1 つ目は，Leader-Follower モデルの導入と Leader から Follower に対する一定の強制力を持った通信を付与することであり，2 つ目は，学習の初期段階において競争的環境にある一方のチームに対しもう一方のチームが「あえて負ける」行動を行うカリキュラム学習と，Train and evaluation による学習フレームワークの工夫である．提案手法の有効性を確認するため，追跡問題に対する学習を複数の手法を用いて実験し，従来からある他の手法と比較して提案手法の有効性を確認するとともに，提案手法を構成する要素毎の影響についても分析がなされた．

次に 2.について，強化学習においてエージェントの方策を特徴づける報酬設計について議論する．複数のエージェントが同時に学習する MARL では，エージェントの自律性を損なうことなく分権的に全体として好ましい協調を実現することが望まれる．複数のエージェントが行った共同行動の結果に対する報酬だけではなく，個々のエージェントの貢献度に応じた報酬信号を設計することができれば，学習エージェントは容易に全体にとって望ましい行動を行うような学習を実現することができる．しかし，協調タスクにおけるインセンティブとしての報酬設計はエージェントの貢献度分配問題に帰着し，協調すべきエージェント達の誘因を損なうことなくシステムが要求する目標を実現するような報酬関数を設計することは容易ではない．一方メカニズムデザインは，ミクロ経済学とゲーム理論の一分野であり，複数の利己的なエージェントをいかにして効率よく取りまとめるかという問題を扱い，社会的余剰が最大となるような設計を行う．本研究では，メカニズムデザインの一例として Vickrey-Clarke-Groves メカニズムによる支払いのルールである迷惑料の考え方に基づいて，個々のエージェントが仮に存在しなかった場合の社会の効用の差分に基づいてそのエージェントに対する報酬を計算する手法を提案し，実験をにより提案手法が評価された．

VCG メカニズムでは，評価対象のエージェントの貢献を評価するためには，そのエージェントが存在しなかった場合の外部性を評価する必要がある．VCG メカニズムに基づく支払いを報酬設計の適用可能性を拡大させるため，エージェントの不在性評価が容易ではない問題であっても適用可能な方法として，評価対象のエージェントが存在しない仮想環境を用いた MARL 手法についても提案する．2 種類の問題設定を対象に学習を行って結果を評価する実験を行い，各種従来手法と比較して議論がなされた．

以上から，本研究では 1.チームを構成するエージェントの組織構造と学習方式に対する工夫，および 2.報酬メカニズムの設計，の 2 点を提案し，従来課題であった環境複雑性による学習困難性の緩和と，学習エージェント間のインセンティブ設計を反映した報酬設計による協調の発現について実例を示した．

これらの成果は，2 篇の査読付き学術論文誌ジャーナル，および，2 件の査読付き国際会議論文として論文が印刷公表されており，論文内容の審査ならびに博士論文公聴会当日の質疑応答により，博士（工学）の学位授与に十分ふさわしい論文であるとの結論に至りました．