CF-008

マルチエージェント強化学習による役割交代型協調行動に関する研究
Role-Alternation Cooperative Behavior by Multi-Agent Reinforcement Learning

林昌吾† 小倉有紀子†† 谷川智洋† 中川聡† 國吉康夫†
Shogo Hayashi Yukiko Ogura Tomohiro Tanikawa Satoshi Nakagawa Yasuo Kuniyoshi

1. はじめに

AI やロボット技術のさらなる発展に伴い、人間とエージェントの相互作用（HAI）がますます重要度を増している。このような人間とエージェントの円滑な協調行動の問題をコーディネーション問題として捉え、特に互いの役割を交代しながら報酬獲得を目指す役割交代型協調行動に注目する。本研究ではこのようなコーディネーション問題の例として、生物学・社会科学における producer-scrounger game (PS-game) [1] というゲームモデルに着目する。このゲームでは 2 人の人が高警戒状態と低警戒状態の 2 つの役割を互いに交代し合うことで両者ともに効率的に報酬を獲得できる。本研究では、このような状況下で巧みな役割交代行動を示すエージェントの実装を目標とし、マルチエージェント強化学習を用いた PS-game シミュレーションを行うことでエージェントの役割交代行動の性能を検証する。

2. 研究の背景

本章では、本研究で議論するコーディネーションの定義や背景を述べ、先行研究についてまとめる。また先行研究に対する本研究の位置付けを示す。

2.1 コーディネーションの定義

この節では、コーディネーションの定義を述べる。Malone らによれば、コーディネーションとは、“Coordination is managing dependencies between activities.” (コーディネーションとは活動の間の依存性を管理することである。) と定義される [2]。自己と 1 人の他者の 2 者間のコーディネーションということであれば、各エージェントは自己の行動と他者の行動との依存性を管理することになる。しかし、他者の潜在的な意思を正しく推定できなければ他者の行動を予測することは困難である。ゆえに、たとえ互いに協調を望んだとしても互いの行動をうまく調整できないコーディネーション問題が発生し得る。

2.2 PS-game

この節では、コーディネーション問題の例として、生物学・社会科学における producer-scrounger game (PS-game) [1] というモデルに注目し、これに対し機械学習的手法でアプローチする方針を述べる。PS-game とは、ヒトを含めた社会性動物の採食行動をモデル化したものである。2 体のエージェントが食料を探索する低警戒（Low-vigilance） の状態をとるか、外敵に備える高警戒（High-vigilance）の状態をとるかを選択することを繰り返す。食料を探索してばかりでは、ある確率で出現する外敵に襲われるが、外敵に備えてばかりでも獲得できる食料が少なくなる。どちらか一方が外敵に備える行動を取れば、もう一方は比較的安全に食料探索ができる。このような状況下で互いが安全かつ効率的に食料を確保していくことを目的とする。

[1] では 2 人の被験者に PS-game を題材とした宝探し課題を行わせることで、人間が協調的なリスク管理を行えるか否かが検証された。図 1 は PS-game のゲームフローを表している。また、図 2 は黒田、亀田の宝探し課題における得点期待値と報酬の設定を示し、図 3 はその宝探し課題を人同士で実施した際の役割交代の結果を示している。

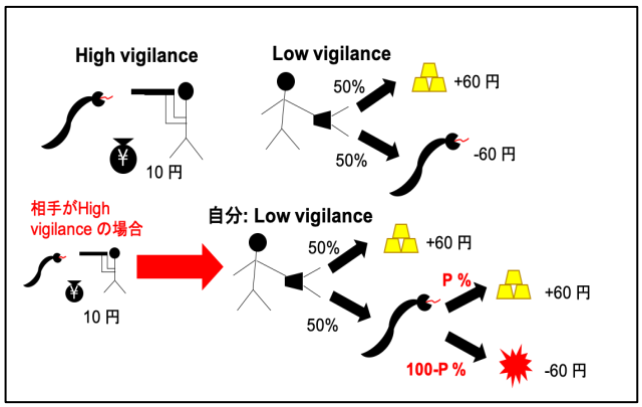

図 1　PS-game ([1]を参考に作成)

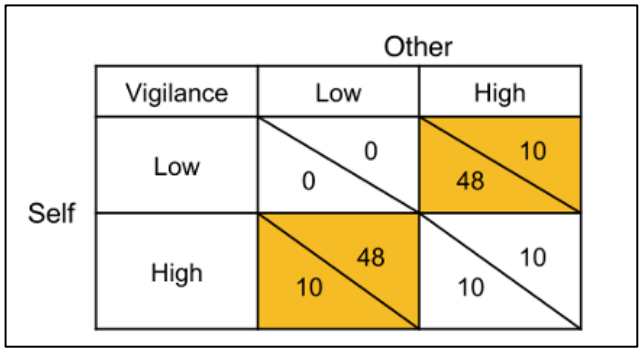

図 2　PS-game における得点期待値表 ([1]より引用)

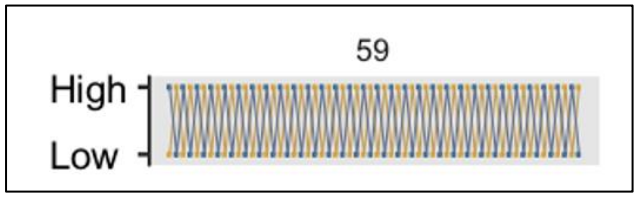

図 3　人同士で PS-game を行った場合の実験結果例 ([1]より引用)

†東京大学情報理工学系研究科 Graduate School of Information Science and Technology, The University of Tokyo
††北海道大学 社会科学実験研究センター Center for Experimental Research in Social Sciences

図 1 に示すように、自分が Low-vigilance で相手が High-vigilance をとっている状態、あるいは自分が High-vigilance で相手が Low-vigilance をとっている状態、この 2 つの状態を繰り返すことで二人が交互に最高報酬を得ることができる。図 3 からは、60 回の試行で 59 回の役割交代に成功し、完璧な役割交代システムを構築することに成功したペアもあったということがわかる。

このゲームにおいて役割交代に着目する理由は次のように説明できる。両者共に High-vigilance 状態をとり続ける場合 (Case1)、 図 1 によると 1 試行ごとに両者平等に reward = 10 を得られるが、これはこのゲームにおいて高い報酬であるとは言えない。

一方、両者共に Low-vigilance 状態をとり続ける場合 (Case2)、1 試行で両者の得られる報酬が reward = 60 となる確率は 50%、reward = −60 となる確率も 50% であるから、結局この場合の両者の報酬期待値は E[reward] = 0 となり、両者 High-vigilance の場合より低い期待値となる。

さらに、一方が Low-vigilance 状態をとり、もう一方が High-vigilance 状態をとる場合 (Case3)、High-vigilance のエージェントはやはり reward = 10 を得る。Low-vigilance のエージェントは、例えば図 1 で P = 80% とした時、1 試行で両者の得られる報酬が reward = 60 となる確率は 50 + 80 / 2 = 90%、reward = −60 となる確率は残りの 10% となる。よってこの場合の Low-vigilance エージェントの報酬期待値は E[reward] = 60 × 90% + (−60) × 10% = 48 となり、これが本ゲーム中では最高の報酬期待値となる。また両者の獲得報酬期待値の平均も (10+48) / 2 = 29 となり最も高くなる。

しかし両エージェントがこの状態を固定的にとり続ける場合、Low-vigilance のエージェントが一方的に High-vigilance のエージェントから利得を搾取し続けることになる。そこで、複数回の試行を行うことを前提とする場合は、互いの Low-vigilance、High-vigilance という 2 つの役割を互いに交代し続けることで、両者が平等かつ効率的に高い報酬を獲得することができる。したがってこのように互いの役割を柔軟に交代することができるかということが問題となる。

本研究では、以上に述べた PS-game のようなコーディネーション問題を解決するために、巧みな協調や役割交代が可能なエージェントの学習方法を、2 体のエージェントを用いたシミュレーションによって機械学習手法の観点から探求する。

2.3 先行研究

この節では、PS-game、役割交代協調行動、マトリックスゲーム等、本研究に関連する先行研究をまとめる。

2.3.1 役割交代型協調行動についての先行研究

[1] にみられるような役割交代型の協調行動についての先行研究についてまとめる。

飯塚、池上は 2 体の車輪付きエージェントの鬼ごっこを力学系でモデリングすることで追う側と追われる側のターンテイキング (役割交代行動) をシミュレートしている [3]、[4]。

Skantze は、会話における予測的かつ連続的なターンテイクのモデルを、Long Short-Term Memory Recurrent Neural Network (RNN) [5] を用いて提示し、人間の場合や先行研究よりも優れた結果を示したことを報告している [6] 。

また、Raffensperger らは、創発的コミュニケーションにおけるターンテイクに関する先行研究を再解釈することによってターンテイクの評価指標を説明し、録音された人間の会話に対してターンテイクの分析を行っている [7] 。

さらに同じく Raffensperger らは、媒体アクセスゲームと呼ぶ一群のステートフルゲームを、人間と機械のコミュニケーションのモデルとして記述し、エージェントの報酬関数に基づいて Q-learning エージェントのターンテイクを予測するために、定常政策を持つエーシジェントのペアによって行われるこれらのゲームのナッシュ均衡を使用する方法を示している [8] 。

2.3.2 マトリックスゲームにおけるマルチエージェント強化学習の先行研究

PS-game はマトリックスゲームの一種として定式化できる。マトリックスゲームとは、各プレイヤーの行動の組によって 図 2.2 のようなペイオフマトリックス (利得行列) で定められた報酬が与えられるようなゲームである。同様にマトリックスゲームとして表現される環境におけるマルチエージェント強化学習の先行研究はいくつかある。

例えば、Sandholm、Critess は、代表的なマトリックスゲームの一つである囚人のジレンマを Q-learning エージェントに繰り返しプレイさせ、固定的な戦略に対して最適な解を導き出すことを示した [9]。

一方 Claus、Boutilier は、簡単なマトリックスゲームで表現できるコーディネーションゲームのゲーム構造と収束性についての議論を展開している [10]。

Leibo らは、囚人のジレンマをはじめとしたマトリックスゲームを社会的ジレンマの一つのモデルとして捉えた [11]。彼らは、現実の社会的ジレンマでは、協力性は「行動」というより「方策」に適用される性質であることに注目し、逐次的社会ジレンマを導入することで共有資源をめぐる競争からどのように紛争が発生しうるかを明らかにしている。

ただしこれらの研究で取り上げられるマトリックスゲームは、各プレイヤーの行動の組によって報酬が一意に定まるものであり、PS-game のように各プレイヤーの行動が確定しても報酬が確率的に変化する環境とは異なっている。

2.4 先行研究に対する本研究の位置付け

先行研究のサーベイから以下のことが確認されている。

- PS-game のように、各プレイヤーの行動の組が定まってもなお確率的に報酬が変化するマトリックスゲームの研究は多くない。
- また、マルチエージェント強化学習を用いた役割交代型協調行動の先行研究も少ない。

そこで、本研究では、確率的に報酬が変化するマトリックスゲームである PS-game において、マルチエージェント強化学習を行った 2 体のエージェントの役割交代行動を調べる。

PS-game は生物学・社会科学の観点で人間のような社会性動物の行動をモデリングしたものであり、マルチエージェント環境においてこれと同様の問題設定を行うことは、エージェントにある種の「社会性」を与えることを意味する。特に本研究では、人間社会の中で必要とされる、巧み

に役割を交替することで平等かつ効率的に報酬獲得とリスク分散を行うことができるシステムを、マルチエージェント強化学習によって構築できるかということに焦点を当てる。

従来手法では一つの最適解を見つけ、それに向かって方策を収束させるものが多い。一方、本研究では最適解が複数個ある、あるいは複数の試行回数で最適解が表現されるような環境であり、このような環境の研究には発展の余地が大いに残されていると考える。

3. PS-game 実験と結果

本章では、本研究で行った実験の詳細、用いた手法、得られた結果について述べる。

3.1 実験 1: Q-learning による実験

実験 1 では、Q-learning [12]を利用して 2 体のエージェントを学習し、PS-game を行った結果を考察する。エージェントは、状態行動価値関数 (Q 関数) を Q-learning アルゴリズムによって更新する。

$$Q(s_t, a) \leftarrow (1-\alpha)Q(s_t, a) + \alpha[r_{t+1} + \gamma \max Q(s_{t+1}, p)]$$

$Q(s_t, a)$ は、エージェントの時刻 t における状態 s_t と行動 a における Q 関数、αは学習率、γ は割引率である。また r_{t+1}とは時刻 $t+1$ においてエージェントが状態 s_{t+1} に遷移した時に得られる報酬である。また、協調行動を促進させるため、エージェントに利他性・公平性を持たせる。ここで、利他性とは、「自己以外のエージェント (他者) がより高い利得を得ることにどの程度積極的な価値を置くか」を示す性質とし、向社会性レベル α の大小によって定義する[13]。α は自己利得と他者利得のどちらにどれほど大きな価値を置くかを表す重みづけ係数であり、エージェントの利他性を定義するに足る性質を有していると考えられる。

次に公平性とは、「自己と他者の報酬の差がより小さくなることにどの程度積極的な価値を置くか」を示す性質であるとし、不公平回避係数 β_{i1}、β_{i1} の大小によって定義する[14]。ここで β_{i1}は他者が自己より大きな報酬を得ることに対する不平等を補正する係数、 β_{i2}は自己が他者より大きな報酬を得ることに対する不平等を補正する係数である。これら利他性、公平性の影響を学習に反映させる。

3.2 実験 1: 結果

実験 1 では、Q-learning を利用して 2 体のエージェントを学習させ、PS-game を行わせる。

本実験における利他性、公平性に関するパラメータは、$\alpha_1 = \alpha_2 = 0.5$ (完全向社会的エージェント)、 $\beta_{11} = \beta_{21} = 5.0$, $\beta_{21} = \beta_{22} = 0.05$ と調整した。

図 4 では、元の [1] における人を対象とした実験と同様に 1 エピソードに 60 回の試行を行い、これを 4000 エピソード程度繰り返した。この図では、0 を Low-vigilance、1 を High-vigilance 状態として、各プレイヤーの状態の遷移を示している。学習結果を見ると、多くの役割交代が成功するケースもあり、元の実験と同様に 60 回の試行のうち 59 回の役割交代が成功したケースもある。しかし、役割交代が全く行われなかったケースも多かったことがわかる。

次に、作業記憶を持つエージェントで役割分化が起こることが示唆されている [18] ことから、我々は、”memory” (記憶容量) に着目することにした。この実験では, memory を「何ステップ分の状態を記憶し、行動判断に利用できるか」と定義した。すなわち Q 関数の変数として用いる過去の状態数を表すものであり, memory を考慮した Q 関数は, $Q(s_{t-m+1}, s_{t-m+2}, \ldots, s_t, a)$と表現される $(m = memory$ としている。

図 5 では、エージェントの memory の値を変化させ memory と役割交替数の関係、memory と 2 人のプレイヤーの報酬の合計の関係をプロットした。図 5 では、memory の値を 1 から 30 の整数値で動かし、1 エピソードに 60 回の試行を各 memory ごとに 1000 エピソード繰り返した。このグラフからは memory と役割交代回数の間には正の相関があることがわかる。一方、memory が大きくなるにつれて総報酬の絶対値が増加する傾向が見られるが、負の方向に大きく落ち込むこともあった。これは memory が増加するほどエージェントが過去の報酬の影響を大きく受け、どの報酬の影響を受けたかによって結果が大きく変化するからであると考える。

また、 “good-alternations episode” と “no-alternations episode” を以下のように定義し、その性質を検討している。 “good-alternations episode” とは、閾値以上の交替が発生したエピソードである。本実験では、1 エピソード 60 回の試行のうち、8 回の交替を閾値とした。これに対して、全く交替が起こらないエピソードを “no-alternations episode” とする。これら good-alternations episode, no-alternations episode を縦軸, memory を 横軸にとってプロットした結果が図 6 である。

good-alternations episode の割合は、memory が非常に小さい時にはほとんど 0% となっている。しかし memory の増加により、最大で 30% 付近まで上昇している。逆に no-alternations episode の割合は、memory が非常に小さい時にはほとんど 100% であるが、memory の増加で最小 20% 程度まで下降している。以上から、memory と good-alternations episode には正の相関関係、memory と no-alternations episode には負の相関関係があることが示唆される。

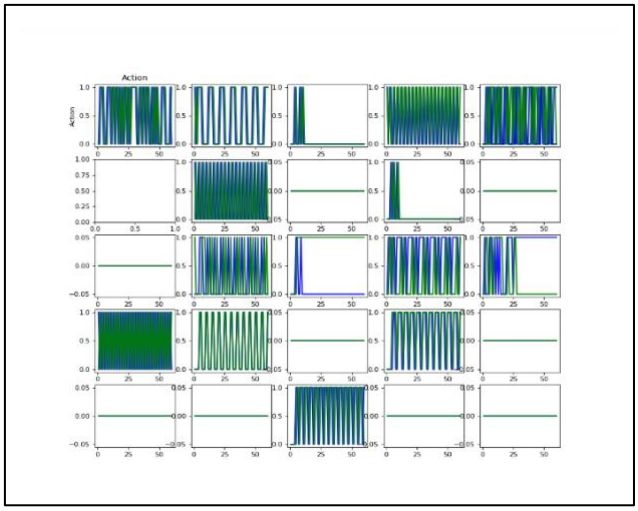

図 4 PS-game 実験結果の一部

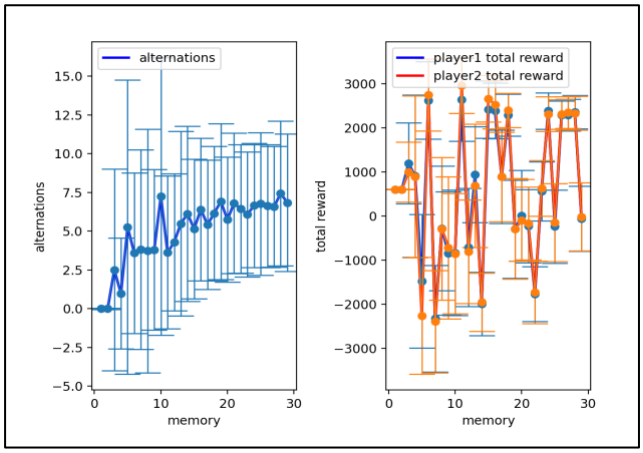

図 5　memory と役割交代回数、合計報酬との関係

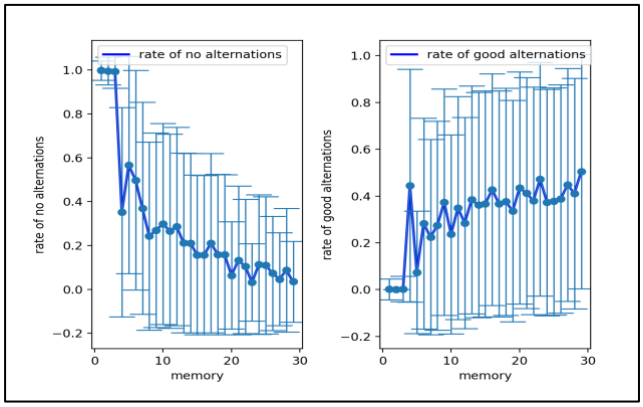

図 6　memory と good-alternations episode、no-alternations episode

3.3 実験 2: 各種マルチエージェント強化学習を用いた実験

実験 2 では DDPG[15]、MADDPG[16]、ROMMEO[17]といった、Q-learning より複雑とされるマルチエージェント強化学習の手法でエージェントを学習させ、PS-game におけるパフォーマンスを検証した。

DDPG (Deep Deterministic Policy Gradient) や MADDPG (Multi-Agent DDPG) は、方策勾配を用いた actor-critic 手法によって様々な状態・行動空間でのオンライン学習を可能にする手法である。この手法では、エージェントの現在の方策を指定する actor 関数 $\mu(s|\theta^{\mu})$および Q 学習の Q 関数にあたる critic 関数 $Q(s,a)$という 2 種類の関数を、ターゲット μ', Q' を目標として学習させる。Critic 関数Qは、ターゲット Q' との差を示す損失関数Lを最小化することを目標とすることで critic を学習させる。一方 actor は、方策勾配 $\nabla_{\theta^{\mu}} J$ を計算することで方策の更新を行う。

ROMMEO (Regularized Opponent Model with Maximum Entropy Objective) は「最適性」を表す確率変数を導入することで、マルチエージェント強化学習を確率的推論として定式化した学習手法である。この手法では、協力的ゲームを「全てのエージェントが協力によって長期的リターンを最大化することができる環境」と定義している。この環境では合理的な相手はある「最適条件」に向かって方策を更新すると仮定しており、この仮定を変分推論によるエージェントの学習アルゴリズムに組み込むことで最適方策や Opponent Model を求められる。

3.4 実験 2: 結果

実験 2 では、DDPG、MADDPG、DDPG-OM (Opponent Modeling)、ROMMEO といった各種マルチエージェント強化学習の手法でエージェントを学習させ、エージェントの平均報酬・方策の推移をプロットした(図 7、図 8、図 9、図 10)。これらの手法は変動環境でオンライン学習が求められるタスクにおいて高いパフォーマンスを示すことが分かっている。実験 2 では実験 1 と異なりエージェントは memory を持たなかったが、 試行回数を重ねることで各エージェントが一定の方策に収束し、役割交代率が 0 に落ち込んでいる。

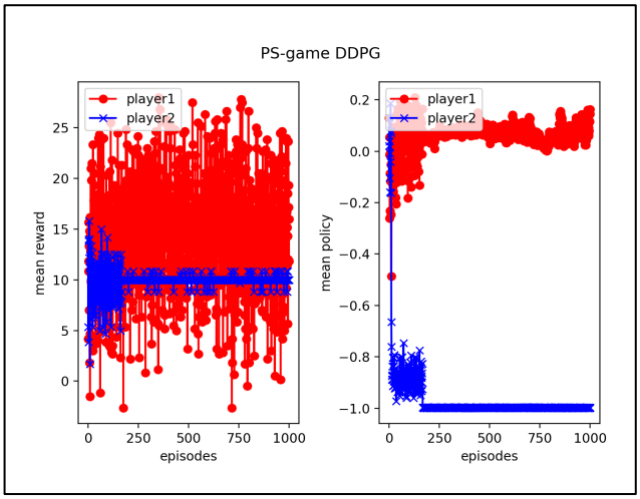

図 7　PS-game DDPG

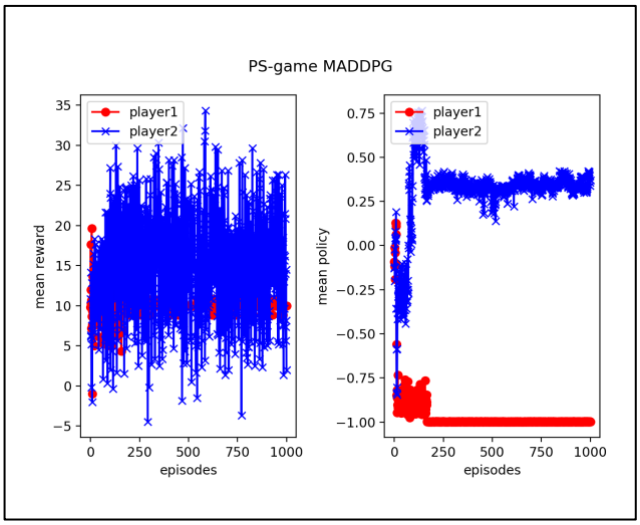

図 8 PS-game MADDPG

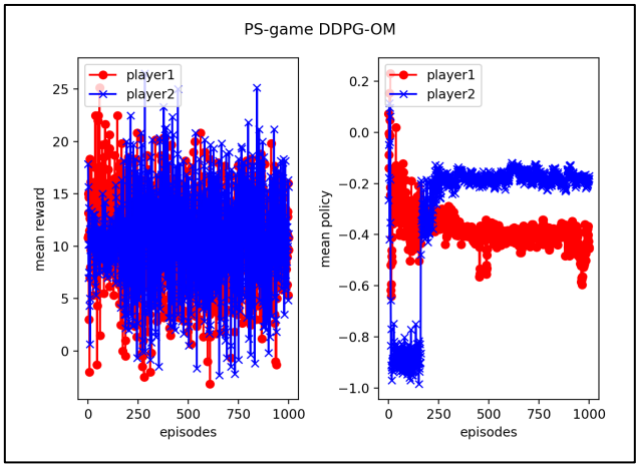

図 9　PS-game DDPG-OM

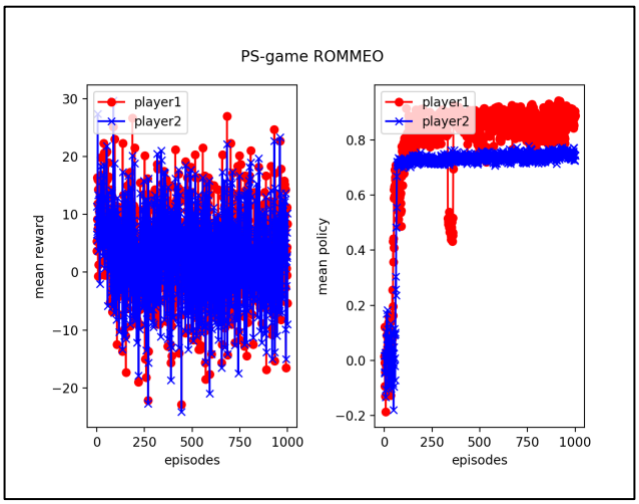

図 10　PS-game ROMMEO

4. 考察

まず、実験 1 の結果をまとめると、以下のようになる。交替回数と good-alternations episode は memory を増やすことで増大させることができ、no-alternations episode は記憶量を増やすことで減少させることができる。そして、このことから、過去の状態を参照する能力が交替を誘発することがわかる。

一方、参照する情報量が多くなったとしても、それらの情報を適切に利用できていなければ性能向上に繋がるとは限らない。本実験では、エージェントは記憶の範囲であれば直前の状態であっても、数十ステップ前の状態であっても並列に扱っているが、これが適切な過去情報の処理のあり方であるかは議論の余地がある。実際の人間等の生物であれば、昔のことよりは、直前に起きた出来事の記憶により強く意思決定を左右されるなどといったことが予想される。

また、実験 2 の結果によると、DDPG、MADDPG、DDPG-OM では、一方のエージェントが High-vigilance 状態、他方が Low-vigilance 状態に収束する結果となっている。これは一方のエージェントが他方に報酬を搾取し続けられる状態を意味している。一方 DDPG-OM、ROMMEO では両エージェント共に High-vigilance 状態に向かっている。これは両者が Low-vigilance 状態となり、負の報酬を得るリスクを回避した結果と解釈できるが、この場合では両エージェントが得る総報酬は大きくならない。各手法では、学習が進行するにつれて役割交代が全く起きなくなっている。これは図 7 や 8 が示すように、エピソードが経過することである方策に収束しているためである。

ここで手法によって収束の様子が異なるのは興味深い結果である。DDPG、MADDPG、DDPG-OM では両者が異なる方策（一方が High-vigilance 状態、他方が Low-vigilance 状態）に収束している。一方 ROMMEO では両者が同じ High-vigilance 状態の方策に漸近しているのが見られる。これについては、ROMMEO の Opponent Model において、相手の方策が High-vigilance 状態をとっていることを読み取った結果、自分も Low-vigilance 状態をとるより High-vigilance 状態をとった方が好ましいという判断を下していると解釈できる。

しかし以上のように、いずれの手法でも学習が進行することにより、役割が次々と交代していくのとは正反対に役割の「固定化」が起こっている。すなわちマルチエージェント強化学習においては Q-learning より高いパフォーマンスを示す可能性がある学習手法を用いたにもかかわらず、memory のないエージェントでは役割交代行動が見られなくなっている。この結果は実験 1 で示唆された役割交代行動における memory の重要性を補強するものともなっている。本実験の設定においては、このような役割の固定化は Low-vigilance 状態をとり続けるエージェントが High-vigilance 状態のエージェントから利得を搾取し続けることを意味する。このような状況を避けて役割交代型協調行動を実現させるためには、今回のようなエピソード数を増やして一定の方策に収束する方法ではなく、柔軟に方策を変更できるモデルが必要と考えられる。

5. おわりに

本研究では、複数のエージェント同士の円滑な協調行動をマルチエージェント環境下のコーディネーション問題として考え、特に役割交代型コーディネーション問題に着目した。そのモデルとして、社会的動物の採食行動をモデルとした PS-game を取り上げた。

実験 1 では、Q-learning エージェントについて、memory を増加させることで役割交替回数と good-alternations episode を増大させることができ、逆に no-alternations episode は減少させることができるという結果を得た。このことから、過去の状態を参照する能力が交替を誘発することが示唆される。

エージェントの意思決定における memory の重要性は、例えば人間の短期記憶と長期記憶の活用に触発された LSTM network [5] 等の手法が時系列データの予測に有効であることからも想像に難くない。ただし、LSTM の場合は学習すべきパラメータが非常に多いことから、学習時間が長い、必要なメモリ量が多いなどの問題点がある。本研

究では比較的少ないパラメータ数を用いたことでより計算負荷を少なくしつつ、memory の活用により役割交代を促す効果が得られたことは注目すべき点である。

一方、実験 2 では、エージェントの役割が固定化する現象が見られ、 Q-learning より複雑な学習手法であっても memory のないエージェントは役割交代をしなくなることがわかった。このことから、相手の方策に応じてより柔軟に方策を変更できるモデルの構築の必要性が浮き彫りになった。

そこで、我々は本研究で扱った PS-game は「リスク管理」の構造を内包していることからリスク感受性 [19] の導入により優れた手法を開発できる可能性を見出している。PS-game において Low-vigilance の状態をとることは、リスクを孕みながらも高い報酬を得られる可能性のある選択である。一方、High-vigilance は確率によらず一定の報酬を得られるリスクの低い選択肢である。このようなリスクの異なる 2 つの方策の選択は、意思決定者が選択肢に対する危険度をどう評価するか、すなわちリスク感受性に依存すると考えられる。人間の場合はリスク感受性に個人差があると考えられ、互いのリスク感受性が特定の条件を満たした場合に役割交代制が創発される可能性がある。

また、ポジティブな記憶の想起が、精神疾患患者のリスク回避性を低下させるという結果が神経経済学や計算精神医学の観点から示唆されている [20] ように、リスク感受性と memory との間にありうる関係というものも興味深い問題であり、さらなる調査が必要である。

PS-game の構造はリスク管理の要素を有する様々な社会的状況に現れることから、本研究が実際の人間社会で優れた役割交代型協調行動を示すロボット・エージェントの開発に貢献することが期待される。本研究の結果は、実際の人間・生物が属する社会における役割制度においても通用する可能性を提示する。

謝辞

本研究は、東京大学次世代知能科学研究センター（AI センター）の支援を受けた。

参考文献

[1] K. Kuroda and T. Kameda, “You watch my back, i’ll watch yours: Emergence of collective risk monitoring through tacit coordination in human social foraging,” Evolution and Human Behavior, pp. 427–48, 2019.

[2] M. Tomasello, A. P. Melis, C. Tennie, E. Wyman, and E. Herrmann, “Two key steps in the evolution of human cooperation: The interdependence hypothesis,” *Current Anthropology*, vol. 53, no. 6, pp.674.692, 2012.

[3] H. Iizuka and T. Ikegami, “Adaptability and diversity in simulated turn-taking behavior,” *Artificial Life (2004) 10 (4)*, pp. 4.61–378, 2004.

[4] T. Ikegami and H. Iizuka, “Turn-taking interaction as a cooperative and co-creative process,” *Infant Behavior and Development 30 (2007)*, pp. 278–288, 2007.

[5] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 178–1780, 1997.

[6] G. Skantze, “Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks,” in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, 2017.

[7] P. A. Raffensperger, R. Y. Webb, P. J. Bones, and A. I. McInnes, “A simple metric for turn- taking in emergent communication,” *Adaptive Behavior 20(2)*, pp. 104–116, 2011.

[8] P. A. Raffensperger, P. J. Bones, A. I. McInnes, and R. Y. Webb, “Rewards for pairs of q-learning agents conducive to turn-taking in medium-access games,” *Adaptive Behavior 20(4)*, pp. 304–4.28, 2012.

[9] T. W. Sandholm and R. H. Crites, “Multiagent reinforcement learning in the iterated prisoner’s dilemma,” *Biosystems*, vol. 37, no. 1, pp. 147–166, 1996.

[10] C. Claus and C. Boutilier, “The dynamics of reinforcement learning in cooperative multiagent systems,” *AAAI/IAAI*, vol. 1998, no. 746-752, p. 2, 1998.

[11] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, “Multi-agent reinforcement learning in sequential social dilemmas,” 2017.

[12] C. J. C. H. Watkins, “Learning from delayed rewards,” 1989.

[13] A. Peysakhovich and A. Lerer, “Prosocial learning agents solve generalized stag hunts better than selfish ones,” 2017.

[14] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Duen˜ez Guzman, A. Garc´ıa Castan˜eda, I. Dunning, T. Zhu, K. McKee, R. Koster, H. Roff, and T. Graepel, “Inequity aversion improves cooperation in intertemporal social dilemmas,” in *Advances in Neural Information Processing Systems*, Eds., vol. 4.2. Curran Associates, Inc., 2018.

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” 2015.

[16] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Neural Information Processing Systems (NIPS), 2017*

[17] Z. Tian, Y. Wen, Z. Gong, F. Punakkath, S. Zou, and J. Wang, “A regularized opponent model with maximum entropy objective,” in *Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 602–608, 2019.

[18] 佐藤尚, 内部英治, 銅谷賢治, et al., “強化学習エージェントによる協調行動とコミュニケーションの創発,” 情報処理学会論文誌数理モデル化と応用 (TOM), vol. 48, no. SIG19 (TOM19), pp. 55–67, 2007.

[19] P. L. Ferreira, F. C. Santos, et al., “Risk sensitivity and theory of mind in human coordination,” PLOS Computational Biology, vol. 17, pp. 1–22, 07 2021.

[20] Shimizu N, Mochizuki Y, Chen C, Hagiwara K, Matsumoto K, Oda Y, Hirotsu M, Okabe E, Matsubara T, Nakagawa S. “The effect of positive autobiographical memory retrieval on decision-making under risk: A computational model-based analysis.” Front Psychiatry. 2022.