様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月　 ７ 日現在

機関番号：１３９０３
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００３０２

研究課題名（和文）仮想社会における強化学習エージェントの報酬評価システム発現過程の解析

研究課題名（英文）Analysis of reward appraisal evolution processes of reinforcement learning agents in a multiagent environment

研究代表者
森山　甲一（Moriyama, Koichi）

名古屋工業大学・工学（系）研究科（研究院）・准教授

研究者番号：１０３６１７７６

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究は，複数の強化学習エージェントが行動する仮想社会における，協力などの社会的な行動の発生に関する研究である．社会的な行動の発生は，比較可能な客観的評価だけでなく，各個体の持つ「価値観」に基づいて行動を学習することで，個体ごとに異なる目的を持つためかもしれない．この考えに基づき，「価値観」が客観的評価に基づいてどのように進化するか，それによりどのような社会が実現するかを計算機シミュレーションと数理的解析で調査した．互いの協調が必要だが，裏切りを選んでしまうジレンマ状況において，エージェントに協調を促す「価値観」が進化すること，および大まかなその進化の方向が明らかになった．

研究成果の学術的意義や社会的意義
強化学習の実現には，状態・行動・報酬の設計が必要である．しかし，複数のエージェントが存在する開いた環境における報酬の設計は非常に困難である．一方で，我々人間は，価値観に基づく主観的な評価（うれしい，恥ずかしいなど）から，複数の人間が存在する開いた社会で適切な振る舞いを学習することができている．本研究は，エージェントの「価値観」の発生・進化を考えることで，開いた環境における報酬の設計を自動化する試みである．同時に，エージェントの「価値観」の形成過程から，人間の価値観などの非合理的側面の存在理由を考える研究でもある．

研究成果の概要（英文）：This research targets the emergence of social behaviors, e.g., cooperation, of reinforcement learning agents in an environment where multiple agents exist. Such social behaviors may emerge if every agent has a different purpose due to learning its behaviors not only from comparable objective evaluation but from its own appraisal. Based on the above discussion, this work investigated how the appraisal system of each agent evolved from the objective evaluation and what society would appear, by computer simulation and mathematical analyses. In a dilemma situation where agents get less payoff by individually rational deception than that by cooperation, we found that the appraisal system evolved to the direction of facilitating cooperation. We also analyzed the direction of the evolution.

研究分野：人工知能

キーワード：知的エージェント　強化学習　報酬設計　進化　マルチエージェントシステム　ゲーム理論

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

強化学習は生物の学習プロセスを模したアルゴリズムであり，エージェントと呼ばれる仮想的な個体の学習を扱う。エージェントは対象とする環境の現在の「状態」を入力として「行動」を出力し，その状態と行動により「報酬」を与えられる。強化学習は，この３つの要素，状態，行動，報酬を結びつけることで，より大きな報酬を得る行動を起こしやすくするアルゴリズムであり，様々なものがこれまでに提案されてきた。しかし，実際に強化学習を用いる際にはこの３つの要素を与えなくてはならず，いずれも今のところ人間が設計している。行動はタスクにおける操作として設計者が定義せざるを得ないが，状態については，近年の深層学習の発展に伴って画像等をそのまま用いることが可能になってきた。さらに，タスクを解決した場合の報奨である報酬については，その設計方法について研究が盛んになってきている。

一方で，我々人間を含む動物は一般に単独で存在するものではなく，複数の個体が同一の環境を共有し，一種の社会を構成している。同様に，複数の強化学習エージェントが仮想社会を形成する場合を考えると，全ての個体が「より大きな報酬」を得ることは困難，あるいは不可能となる。このような環境において，譲り合いなどの協調行動が強化学習により獲得できるかどうか，さらに，具体的にどのようにすればよいかは大きな研究課題である。

２．研究の目的

本研究では，複数のエージェントが個々に意思決定する仮想社会において，個々のエージェントが強化学習により協調行動を学習することを考える。強化学習は生物の学習を模したアルゴリズムであるため，我々人間も用いている学習手法であると言える。しかし同時に，我々は社会において協調することが可能である。それは，行動の結果として得られた利益をそのまま受け入れるのではなく，その行動が適切であったかなどの評価を行っているためと考えられる。そこで，似た仕組みをエージェントに持たせることで，その評価を用いた強化学習により協調行動を学習するエージェントを構築する。具体的には，評価に影響する「価値観」のようなものを持つエージェントを実現することを考える。

しかし，人間の価値観を正当にモデル化してエージェントに実装可能かどうかは明らかではない。そのため本研究では，価値観を所与のモデルとして実装するのではなく，複数のエージェントの相互作用から社会的行動の必要に応じて現れるものと考える。すなわち，本研究は，複数のエージェントからなる仮想社会で他者と相互作用をするエージェントが，自らの評価（主観的評価）に基づいて行動を強化学習しながら，この主観的評価ではなく客観的な利益に基づいて評価システムを進化させることを繰り返すと，どのような評価システムが出現し，その結果としてどのような社会が実現するかをシミュレーションおよび数理的解析により明らかにする。

３．研究の方法

本研究で対象とする仮想社会は，ゲーム理論で人間社会のモデルとして用いられる「ゲーム」，特にその中でもっとも単純な２人２行動同時手番対称ゲームとする。本研究では特に，２人２行動同時手番対称ゲームのうち，もっとも有名で興味深い囚人のジレンマゲームに着目した。このゲームは，個々が自らの利得を大きくする最適な行動（裏切り）を選択すると，結果として得られる利得が相互に協調した場合よりも小さくなるというジレンマをもたらす。エージェントの持つ評価システムを報酬から主観的評価をもたらす関数と考えると，本研究は，この関数が利益に基づく進化によりどのようなものになるのか，どのような過程を経てそのようになるのか，を計算機シミュレーションと数理的解析により調査した。

まず，学習・進化過程の数理的解析の前段階として，計算機シミュレーションによりどのような学習・進化過程が見られるかを調査した。２人２行動ゲーム，強化学習，進化計算アルゴリズムを計算機上に実装し，囚人のジレンマゲームの報酬から主観的評価を導く関数を考案・実装した。それから，様々な評価関数を持つ多数のエージェントを集めた仮想社会を想定し，様々な条件下でシミュレーション実験を多数行うことで，エージェントの学習過程や評価関数の進化過程を調査した。重要な点として，進化の適合度は主観的評価でなく，エージェントの獲得した客観的な利益とした。なぜならば，進化の大前提として，個体として生き残ることが必要であるため，外部から得られる客観的な利益（食料など）により進化が制御されることが自然だからである。このシミュレーションでは，各エージェントが持つ主観的評価関数，強化学習により獲得した方策と選択した行動，獲得報酬，各世代間の主観的評価関数の進化の推移，などを記録した。また，多数のシミュレーションを実行する必要から，その高速化についても検討した。

次に，主観的評価をもたらす関数からなる空間を数理的に考察した。この空間をシミュレーションの出力である仮想社会，すなわち属するエージェントがとる行動のパターンで分類することで，どのような関数がどのような行動をもたらす傾向があるのかを明らかにした。さらに，分割された空間の間を，利益に基づく進化によりどのように遷移するかを調べることにより，主観的評価がどのように誘導されるのかを調査した。

本研究は用いる強化学習手法に依存するため，もともとその手法がどのような行動をもたら

すかなど，手法自体の性質を明らかにする必要がある。そのため，強化学習手法そのものの性質についてもいくつか調査を行った。

## ４．研究成果

本研究では，まず，主観的評価$u$を報酬$r$の３次関数$u(r) = ar^3 + br^2 + cr + d$で表現し，その各項の係数を遺伝子として染色体を構成，遺伝的アルゴリズムを用いて 10000 世代進化させた。このシミュレーション実験を 100 回行ったところ，79 回で繰り返し囚人のジレンマにおける行動選択１回あたりの平均利得が 2.5 $(= (T + S)/2)$ を上回り，少なくとも１回は相互協調が発生することが確認された。$u(r)$に強く影響する３次および２次の係数$a, b$について最終結果を図示すると，図 1 のようにこれらが一定の傾向に収束することが判明した。以下についても，同様に$a$-$b$平面で議論する。なお，シミュレーションの高速化に関して，GPU による並列計算を用いることにより，単一 CPU コアでの実行と比べて約 60 倍の高速化に成功した。

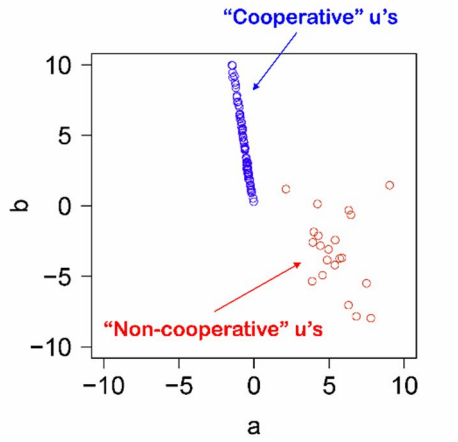


**図１　10000 世代の平均染色体（$a$-$b$平面に射影）**

続いて，主観的評価の進化過程について分析を行った。ある行動によって得られる評価がその他の行動によって得られる評価よりも大きい場合，強化学習により前者の行動を選好すると考えられる。そこで，まず，囚人のジレンマゲームにおける主観的評価から自身の行動のみに依存する評価を定義し，それに基づいてエージェントがどの行動を選好するかを表す指標を行動傾向として定義した。

まず，第一の行動傾向として協調傾向を定義した。自身の行動を$X \in \{C, D\}$としたとき，関数$u_{cp}(X)$および$m_{cp}$を以下の通り定義する。$u_{cp}(X)$は相手が行動を等確率で選択するという仮定の下での各行動の主観的評価を表し，$m_{cp}$は両行動による主観的評価の関係を表す。

$$u_{cp}(X) \equiv \begin{cases} \dfrac{u(R) + u(S)}{2} & \text{if } X = C, \\ \dfrac{u(T) + u(P)}{2} & \text{otherwise,} \end{cases} \qquad m_{cp} \equiv \begin{cases} \dfrac{u_{cp}(C)}{u_{cp}(D)} & \text{if } u_{cp}(D) \neq 0, \\ u_{cp}(C) & \text{otherwise.} \end{cases}$$

これらを用いると，エージェントの協調傾向，すなわちどの程度協調的か（協調を選好するか），裏切り的か（裏切りを選好するか）を$m_{cp}$で表すことが可能になる。

次に，第二の行動傾向として一致傾向を定義した。「エージェントと相手の行動が同一である」という言明を$I$としたとき，関数$u_{cf}$および$m_{cf}$を以下の通り定義する。

$$u_{cf}(I) \equiv \frac{u(R) + u(P)}{2}, \qquad u_{cf}(\neg I) \equiv \frac{u(T) + u(S)}{2}, \qquad m_{cf} \equiv \begin{cases} \dfrac{u_{cf}(I)}{u_{cf}(\neg I)} & \text{if } u_{cf}(\neg I) \neq 0, \\ u_{cf}(I) & \text{otherwise.} \end{cases}$$

これらを用いて，エージェントの一致傾向，すなわち相手と同一の行動あるいは異なる行動をどのくらいの強さで選好するかを$m_{cf}$で表すことが可能になる。

これらの行動傾向を用いると，主観的評価のパラメータ空間を分割することが可能になる。協調選好と一致選好，裏切り選好と不一致選好が同時に現れることが多いが，それぞれの強度を考えると，図 2 のように空間が４つに分割される。

次に，各空間内における進化の方向の説明を試みた。その結果，以下の知見を得ることができた。これらの知見をまとめると，進化が図 2 の D-A 領域と Cf-Cp 領域の境界，すなわち$a$-$b$平面の上部中央付近に向かうことが予想でき，この結果は図 1 の実験結果と一致する。

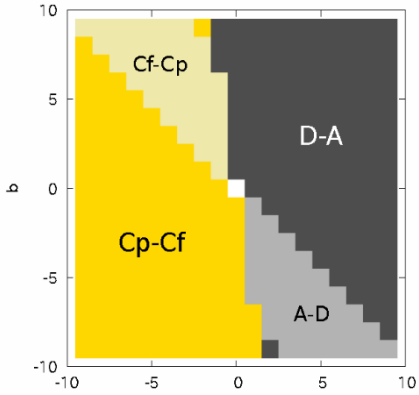


**図２　パラメータ空間の分割**

- 裏切り選好かつ不一致選好で前者が優位な場合（D-A 領域），$m_{cp}$と$m_{cf}$が大きくなるほど協調と一致を選好するようになる。この方向は，$a$-$b$平面上で右下から右上，さらに左上に進む方向になる。協調かつ一致，すなわち相互協調の際の利得が相互裏切りより大きいため，偶然発生した相互協調をもたらす染色体が生き残り，徐々にこの方向に進化し，結果としてこの領域の左上部に向かうとみられる。
- 裏切り選好かつ不一致選好で後者が優位な場合（A-D 領域），$m_{cp}$と$m_{cf}$が大きくなるほど協調と一致を選好するようになる。これは，D-A 領域に向かう方向となる。D-A 領域と

同様に，偶然発生した相互協調をもたらす染色体が生き残り，結果として D-A 領域へ進化が進むとみられる。

- 協調選好かつ一致選好で前者が優位な場合（Cp-Cf 領域），上記とは逆に$m_{cp}$と$m_{cf}$が大きくなるほど裏切りと不一致を選好するようになる。これは A-D 領域に進む方向となる。この領域では協調が優位なため，裏切りの選択が自らの利得を大きくする。したがってこの方向に進化が進み，結果として右に向かうとみられる。
- 協調選好かつ一致選好で後者が優位な場合（Cf-Cp 領域），Cp-Cf 領域と同様に$m_{cp}$と$m_{cf}$が大きくなるほど裏切りと不一致を選好するようになる。一方で，D-A 領域に進む方向ではこれらの値が小さくなり，協調と一致をさらに選好するようになる。Cp-Cf 領域とは異なり，この領域では一致選好が優位なため，一方的な裏切りは好まれず，結果として相互協調を維持して D-A 領域に向けて進化が進むと考えられる。

さらに，本研究は主観的評価を導入することにより，結果としてゲームの利得を（個々のエージェントごとに）変更していることになる。そこで，繰り返し２人２行動同時手番ゲームにおける利得と強化学習のパラメータおよび現れる行動の関係について解析し，囚人のジレンマゲームの条件を満たしていながら，強化学習エージェントの相互協調が発生する割合について数学的に明らかにした。

## ５．主な発表論文等

〔雑誌論文〕（計３件）


(1) Kazuhiro Murakami, Koichi Moriyama, Atsuko Mutoh, Tohgoroh Matsui, and Nobuhiro Inuzuka: Accelerating Deep Q Network by Weighting Experiences. *Proceedings of the 25th International Conference on Neural Information Processing (ICONIP)*, Vol. 1, pp. 204–213, LNCS 11301, Springer, 2018. doi:10.1007/978-3-030-04167-0_19 （査読有）

(2) Masaya Miyawaki, Koichi Moriyama, Atsuko Mutoh, Tohgoroh Matsui, and Nobuhiro Inuzuka: Evolution Direction of Reward Appraisal in Reinforcement Learning Agents. *Proceedings of the 12th KES International Conference on Agent and Multi-agent Systems: Technologies and Applications (KES-AMSTA)*, pp. 13–22, Springer, 2018. doi:10.1007/978-3-319-92031-3_2 （査読有）

(3) Koichi Moriyama, Kaori Nakase, Atsuko Mutoh, and Nobuhiro Inuzuka: The Resilience of Cooperation in a Dilemma Game Played by Reinforcement Learning Agents. *Proceedings of the 2nd IEEE International Conference on Agents (ICA)*, pp. 33–39, IEEE, 2017. doi:10.1109/AGENTS.2017.8015297 （査読有）


〔学会発表〕（計４件）


(1) 千賀喜貴, 森山甲一, 武藤敦子, 松井藤五郎, 犬塚信博: GPGPU を用いた強化学習エージェントの並列進化シミュレーション, 人工知能学会全国大会（第 32 回）, 2018.

(2) 村上知優, 森山甲一, 武藤敦子, 松井藤五郎, 犬塚信博: 経験データ重み付けによる Deep Q Network の高速化, 人工知能学会全国大会（第 32 回）, 2018.

(3) 宮脇昌哉, 森山甲一, 武藤敦子, 松井藤五郎, 犬塚信博: マルチエージェント強化学習における主観的効用の進化過程に関する分析. 人工知能学会全国大会（第 31 回）, 2017.

(4) 黒木是冶, 森山甲一, 武藤敦子, 松井藤五郎, 犬塚信博: GPGPU を用いた 2 人ゲームにおける強化学習の高速化. 情報処理学会第 79 回全国大会, 2017.


## ６．研究組織

(1) 研究分担者　なし