原著論文

強化学習と情動学習に基づく意思決定法： 利己的な判断による協調行動の創発[†]

三澤 秀明*・松田 充史*・堀尾 恵一*

情動は、動物が生存していく上で非常に重要な要素であり、社会的な意思決定にも関係していると考えられている。本論文では、恐怖情動に基づく利己的な判断による社会的行動の創発を検証するために、強化学習と情動学習によって学習した価値に基づいて行動選択を行う意志決定法を提案する。強化学習ではタスクを達成するための行動規範を学習し、情動学習では環境からの入力に対する情動的な価値を学習する。自身に対する不快を回避するという情動の割り込み機能を模擬した意思決定では、強化学習と情動学習によって得られた価値を用いて、他のエージェントや状況に合わせて自身の行動を調整する。目標探索問題のシミュレーション実験の結果、提案手法を用いたエージェントが社会的行動である協調行動を創発すること、環境変化への適応性を有することを確認した。

キーワード：情動、強化学習、情動学習、協調行動、意思決定

1. はじめに

ロボット技術の進展によって、産業用ロボットに加え、人間の生活環境内で作業を行うロボットが精力的に開発されている[1]。自律型移動ロボットによるサッカーで有名なロボカップでは、人間とロボットの共存を目指し、キッチンやリビングルームなどの家庭環境において、ロボットが様々な課題に取り組むロボカップ@ホームという競技も行われている[2]。このようなロボットには、少子高齢化社会における労働力として、多くのサービス分野での活躍が期待されている。人間と同じ生活環境で共生するロボットには、自身の置かれた状況だけではなく、人間や他のロボットとの関係を考慮した社会的な意思決定が求められる。

本研究では、社会的な意思決定は情動の機能によっても発現されるという考えに基づき、エージェントの意志決定に情動の概念を利用する。心理学の分野では、外部刺激や観念によって生じる快-不快の意識現象である感情のうち、特に一過性に生じる強い感情を情動と呼ぶ[3]。近年の脳科学の進歩によって、動物の情動に関係する神経回路メカニズムが明らかにされてきた。特に、大脳辺縁系に属する扁桃体は、恐怖情動に関する古典的条件づけに深く関係している[4]。

恐怖情動は、動物が生存していく上で非常に重要な要素であり、自身に対する危険や不快な状況を回避する際に発現される。扁桃体は、環境からの入力に対して、これまでの経験に基づいて、快-不快といった情動的な価値を評価する。この評価に応じた扁桃体からの出力が、他の脳領域に送られ、すみ行動や血圧の上昇などの情動反応が引き起こされる。古典的条件づけによって学習された情動的な価値は、オペラント条件づけで学習した行動に関する意思決定に対しても影響を与えられている[5]。扁桃体を損傷すると、他者の表情が認識できなくなったり、社会的行動が取れなくなったりすることから、扁桃体は他者の意図や情動を理解する社会的な認知機能や社会的な意思決定においても、重要な役割を果たしていると考えられている[6]。

情動の概念は、これまでも、エージェントシステムや自律ロボットへ応用されている。石黒らは、情動機構を工学的にモデル化し、そのモデルを用いた自律移動ロボットのための新たな行動調停法を提案した[7]。提案された情動機構では、ロボットの感情を単純に快-不快に分類し、それをフラストレーションレベルという変数により表現している。このフラストレーションレベルをもとに、目標点接近行動と障害物回避行動の調停を実現している。瀧田らは、動物における積極性と警戒心に対応するような形で情動の概念を利用して価値判断の変更を行う強化学習手法を提案した[8]。提案されたモデルは、報酬の最大化のみを目的とした積極的モジュールと、罰の最小化のみを目的とした警戒的モジュールから構成される。それぞれ

[†] Decision Making Based on Reinforcement Learning and Emotion Learning: Emergence of Cooperative Behavior with Selfish Judgment

Hideaki MISAWA, Atsushi MATSUDA and Keiichi HORIO

* 九州工業大学大学院生命体工学研究科

Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology

のモジュールで学習された評価値が、バイアスパラメータによって重みづけされた形で統合され、行動選択が行われる。餌探索シミュレーションによって、積極性と警戒心の動的な調整によって餌探索行動の学習精度が向上することを示している。久保田らは、ロボットが環境や人間との適応的な相互作用を実現するための情動モデルを提案した[9]。提案された情動モデルでは、外界からセンサによってロボットが知覚した情報を用いて、ロボットの内部状態である情動パラメータ、感情パラメータ、雰囲気パラメータを相互に変化させる。この情動モデルと強化学習、ファジィ推論を用いて、障害物回避などの行動選択を実現している。これらの研究では、情動の概念を意思決定に取り入れているが、複数のエージェントが存在するマルチエージェント環境における社会的行動と情動との関係については述べられていない。

ロボットなどの複数のエージェントによる社会的行動の実現は、マルチエージェントシステムにおける行動学習として研究されてきた。高玉は、各エージェントが独自の目的を遂行するために独立して学習する分散学習を個体レベルでの適応、エージェント間で情報を共有する協調学習をグループレベルでの適応と分類している[10]。行動学習手法としては、多くの場合、環境との相互作用によって、その環境における最適な行動規範を獲得できる強化学習が用いられる[11]。森山らは、マルチエージェント環境における社会的ジレンマの問題に着目し、環境の状況に応じて自律的に自己の利益を放棄することにより、社会的ジレンマを解決する強化学習手法を提案した[12]。社会的ジレンマとは、個人の利益追求と集団における利益追求が合致しておらず、個人の利益追求を行えば集団における利益追求が阻害され、集団における利益追求を行えば個人の利益追求が阻害される葛藤(ジレンマ)のことを指す。提案手法では、自身および近隣他者の現在の報酬和を学習用報酬として用いることによって、社会的ジレンマを回避する手法を提案している。森山らの手法は、エージェント間で報酬情報を共有するので、グループレベルでの適応と考えられる。柴田らは、マルチエージェント環境において、利害の衝突回避のために個性と社会性が発現するという考えを提唱し、強化学習による行動学習によって個性や社会性が発現するかについての検証を行った[13]。柴田らは、個性を“感覚入力から行動へのマッピングが個体ごとに異なること”、社会性を“感覚入力から行動へのマッピングは個体間で同じであるが、感覚入力の違いによってのみエージェントの行動が異なること”と定義している。電車乗り降りシミュレーションにおいて、個々の

エージェントが、自らが獲得した報酬に基づいて強化学習を行うことによって、個性と社会性の分化が発現することを確認している。柴田らの手法では、各エージェントが利己的に行動する状況下で、環境の違いなどの要因によって社会的行動が創発されたので、個体レベルでの適応であると考えられる。これらの手法は、マルチエージェント環境における社会的行動を取り扱っているが、情動の概念を用いていない。

本研究では、柴田らと同様に、利己的な判断による社会的行動の創発を考える。ただし、自身に対する不快を回避するという情動に基づいた利己的な判断を取扱う。例えば、チーム作業を行う場合、一人の作業遅延が全体のスケジュールの遅延につながり、結果として、そのチームの評価が悪くなる場合がある。このような場合、チームのメンバーが、その一人の作業支援を行うことが考えられる。この支援は、単純に相手のことを思いやり、援助する利他的な意志決定による支援だけでなく、そのチームに所属する自身の評価を下げないために行う利己的な意志決定による支援であることも考えられる。つまり、自分に悪い評価が与えられるという自身に対する不快な状況を避けるために、他者を支援するという社会的な意志決定が行われていることが考えられる。この自身に対する不快を避ける利己的な判断を、恐怖情動の機能としてエージェントの意志決定に組み込むことによって、情動と社会的行動の創発の関係性を検証する。

本論文では、強化学習と情動学習に基づいて行動選択を行う意志決定法(Decision making based on Reinforcement learning and Emotion learning, DRE)を提案する。DREを用いたエージェントは、環境における最適な行動を学習しつつ、自己と他者との関係性における情動経験を学習する。本研究では、情動経験に基づき、環境からの入力に対する情動的な価値(情動価値)を学習することを情動学習と呼ぶ。これらの学習によって得られた価値を利用することで、エージェントは、他のエージェントの行動や状況に合わせて自身の意志決定を調整する。シミュレーション実験では、DREを用いたエージェントが社会的行動を創発することができることを確認し、DREの有効性を検証する。

2. 強化学習と情動学習に基づいた意思決定

2.1 問題の枠組み

本論文では、DREの適用対象として、複数のエージェントと複数の目標が存在する状況下での目標探索問題を考える。エージェントのタスクは、行動制限回

数内に目標の1つに、他のエージェントよりも先に到達することである。タスクとは別に、エージェントは、ある条件が満たされたときに、恐怖条件づけのための罰を受ける。罰を引き起こす条件としては、「エージェントが社会的行動を取らない」とする。しかしながら、エージェントは、罰を受ける条件を知らない。エージェントには、この罰が恐怖情動の経験となる。エージェントは、タスクを達成するだけでなく、罰を受ける危険を回避する必要がある。各エージェントが、(1)自身への危険回避と(2)報酬を得るためのタスク達成という2つの利己的な行動を取ることが、エージェントの社会的行動の創発につながるかを検証する。

2.2 DRE

情動は、個体を何らかの行為に駆り立てる強い動機づけの性質を有する[14]。この特性は、ある特定の事象や対象に意識を没頭させる働きをする。つまり、情動は、進行中の活動がいかなるものであっても強引に割り込み、その活動を否定なく分断する割り込み機能として働く。このような情動の割り込み機能は、天敵に遭遇するなどといった危機的な状況において、動物が生き残るために合理的な働きを行っていると考えられる。

本研究では、自身に対する不快を回避するという恐怖情動による割り込み機能を実現するDREを提案する。図1に、DREを用いたエージェントの構成を示す。DREを用いたエージェントは、強化学習モジュール、情動学習モジュール、行動決定モジュールの3つのモジュールから構成される。強化学習モジュールでは、エージェントがタスクを達成するための状態価値を強化学習によって学習する。情動学習モジュールでは、エージェントが罰を受ける危険を回避するための情動価値を情動学習によって学習する。行

動決定モジュールでは、タスク達成のための状態価値による行動選択に、罰(恐怖体験)を回避するための情動価値による判断を加え、情動による割り込み機能を実現する。強化学習は、ある行動の結果によって生じる報酬や罰によって、その行動の生起頻度が変化するオペラント条件づけを実現するための方法の1つであると考えられる。情動学習では、恐怖情動を経験した状況を表す情動状態に対する情動価値を学習する。罰を経験する前には、エージェントはその状況に対しては恐怖を感じないが、学習後にはその状況に対して恐怖を感じるようになる。この情動学習は、中性的刺激に対する情動反応を学習する古典的条件づけに相当すると考えられる。提案するDREでは、この2つの条件づけに基づく意思決定によって、情動の割り込み機能を実現する。以下では、各モジュールの詳細について説明する。

2.2.1 強化学習モジュール

強化学習は、エージェントが環境における累積報酬を最大化する行動規範を、試行錯誤によって獲得するための手法である[11]。強化学習モジュールでは、タスクを達成するための行動選択に必要な状態価値を学習する。本論文では、学習する価値の数を少なくするために、行動価値でなく状態価値を用いる。状態価値の学習には、環境モデルを必要とせず、経験から直接学習を行うモンテカルロ法を用いる[11]。なお、今回の問題設定では、モンテカルロ法によって状態価値を学習する方法を採用したが、問題設定によっては、この方法では十分に学習できない場合もあると考えられる。そのような場合には、Temporal Difference 学習、Q学習、Actor-Critic 法など、他の強化学習手法を用いる必要があると考えられる。

エージェントが行動制限回数内に目標に到達することができた場合には、報酬 R がエージェントに与えられ、状態価値 $V(s_t)$ ($t=1, 2, \dots, t'$) が次式によって更新される。

$$V(s_t) \leftarrow V(s_t) + \alpha \cdot (r_t - V(s_t)) \quad (1)$$

$$r_t = \gamma^{t'-t} \cdot R \quad (2)$$

ここで、 α は学習率、 r_t は行動ステップ t に通過した状態に与えられる報酬、 t' は1エピソードにおける最終行動回数、 γ は減衰率を表す。本研究では、エージェントがエピソード中に同じ状態を2回以上訪れた場合、最後の訪問のみを更新の対象とする。エージェントが目標に到達できなかった場合には、環境の変化に対応するために、次式によって全状態の価値を減衰させる。

図1 DREを用いたエージェント

$$V(s_t) \leftarrow \eta \cdot V(s_t) \quad (3)$$

ここで、 η は減衰率である。

2.2.2 情動学習モジュール

情動学習モジュールでは、エージェントが経験した恐怖情動を情動価値として学習する。情動価値 $E(u_t)$ は、恐怖情動を経験した状況を表す情動状態 u_t に対して定義される。社会的行動は、2つ以上のエージェントが存在する状況で、他のエージェントとの関係を考慮した上で、取られる行動である。したがって、情動状態 u_t は、エージェント間の関係性を表現する必要がある。本研究では、各エージェントが存在する位置の組み合わせを情動状態とする。情動価値 $E(u_t)$ の学習には、モンテカルロ法を用いる。強化学習モジュールと同様に、問題設定によっては、他の強化学習手法を用いる必要があると考えられる。

エージェントが社会的行動を取らずに罰 D を受けた場合には、情動価値 $E(u_t)$ ($t = 1, 2, \dots, t^*$) は次式によって更新される。

$$E(u_t) \leftarrow E(u_t) + \alpha' \cdot (d_t - E(u_t)) \quad (4)$$

$$d_t = \zeta^{t^* - t} \cdot D \quad (5)$$

ここで、 α は学習率、 d_t は行動ステップ t に経験した情動状態に与えられる罰、 t^* は罰を受けた行動ステップ、 ζ は減衰率を表す。罰を受けたエピソード中に経験しなかった情動状態の価値は、最新の情動経験を優先させるために、次式によって情動価値を減衰させる。

$$E(u) \leftarrow \beta \cdot E(u) \quad \text{for } u \notin U_d \quad (6)$$

ここで、 β は減衰率、 U_d は罰を受けたエピソード中に経験した情動状態の集合である。罰を受けなかったエピソードでは、すべての情動価値はそのまま保持される。

2.2.3 行動決定モジュール

行動決定モジュールでは、強化学習モジュールと情動学習モジュールで学習した価値をもとに、行動を決定する。行動決定モジュールは、行動選択プロセスと行動判断プロセスの2つのプロセスから構成される。両方のプロセスにおいて、 ε -グリーディー法を用いる。

行動選択プロセスでは、エージェントがタスクを達成するための行動を選択する。 ε_s の確率でランダムに行動を選択し、 $1 - \varepsilon_s$ の確率で状態価値に基づいて行動を選択する。状態価値に基づいた選択では、エージェントの周辺で最も状態価値が高い方向への行動を選択する。

行動判断プロセスでは、エージェントが過去に経験した恐怖を避けるために、行動選択プロセスで選択された行動に、情動に基づいた判断を加える。 ε_e の確率で行動選択プロセスにおいて選ばれた行動をそのまま次の行動と判断し、 $1 - \varepsilon_e$ の確率で情動価値に基づいて行動を判断する。情動価値に基づいた判断では、他のエージェントの現在の位置を基準として、エージェントの周辺で最も情動価値が高い方向への行動を選択する。以上をまとめると、 $\varepsilon_s \cdot \varepsilon_e$ の確率でランダムに、 $(1 - \varepsilon_s) \cdot \varepsilon_e$ の確率で貪欲に、 $1 - \varepsilon_e$ の確率で情動に基づいて行動決定が行われる。 ε_e として小さな値を用いれば、情動価値に基づく判断が優先され、情動による割り込み機能が実現される。なお、DREに関する記述は、目標探索問題という問題設定に依存しており、一般的な問題を対象とした記述にはなっていないことに注意されたい。提案手法における行動決定モジュールでは、タスク達成のための状態価値による行動選択に、罰を回避するための情動価値による判断を加えるという方法で、情動による割り込み機能を実現している。情動の割り込み機能として別の実現方法を用いる場合には、Actor-Critic 法など、他の強化学習手法を用いることも可能であると考えられる。

3. シミュレーション実験

3.1 実験設定

本論文では、2つのエージェントと2つの目標が存在する状況での目標探索問題を取り上げる。図2に、シミュレーションで用いた離散グリッド空間を示す。

3.1.1 DREエージェント

DREを用いたエージェント(DREエージェント)のタスクは、行動制限回数内に、目標の1つに他のエージェントよりも先に到達することである。タスクを達成したエージェントには、報酬 R が与えられる。

図2 探索グリッド空間 (15×15). (a) 実験1における目標位置. (b) 実験2における1000エピソード後の目標位置.

エージェントが社会的行動を取らなかった場合には、罰 D が与えられる。実験では、「各エージェントが異なる目標に到達する行動」を協調行動と呼び、この協調行動を社会的行動と考えた。行動制限回数内に協調行動を達成できなかった場合に、両方のエージェントに罰 D が与えられる。エピソード中に衝突が起きた場合にも、両方のエージェントに罰 D が与えられる。実験では、エージェントが同じ位置に移動することを衝突と定義した。なお、両方のエージェントが目標に到達しない場合は、協調行動を達成できなかった場合に含まれる。

強化学習モジュールでは、状態 s として、図2のセルを用いた。情動学習モジュールでは、情動状態 u として、探索グリッド空間を 5×5 に分割したエージェントエリアの組み合わせを用いた。図3にエージェントエリアと情動状態の関係を示す。エージェントエリアは、 3×3 のセルから構成される。全情動状態を表す情動マップは、 $625 (25 \times 25)$ 状態から構成される。各エージェントは、他のエージェントが存在するエ

アを観測できるものとした。セルではなくエリアを用いた理由は、(1)状態数を少なくするためと、(2)動物がある地点で過去に恐怖経験をした場合、その地点だけでなく、その周辺を避けると考えられるためである。行動選択プロセスにおける状態価値に基づく選択では、エージェントの周辺で最も状態価値が高いセルへ移動できる行動を選択した。行動判断プロセスにおける情動価値に基づく判断では、他のエージェントのエリアを基準として、エージェントの周辺で最も情動価値が高いエリアへ近づくことができる行動を選択した。状態価値が最も高いセルや情動価値が最も高いエリアが複数ある場合には、その中からランダムに選択した。

3.1.2 比較対象

各実験では、DREエージェントの比較対象として、以下の3つのエージェントを用いた。

1つ目の比較対象は、強化学習のみを行うエージェント (RLエージェント) である。RLエージェントのタスクは、DREエージェントのタスクと同じである。RLエージェントは、情動学習モジュールを除いたDREエージェントと同じであり、 ε_s のみを用いて行動決定を行う。RLエージェントにおいては、タスクを達成したエージェントに報酬 R が与えられ、罰 D は用いない。

他の2つの比較対象は、協調行動を目的とした強化学習を行うエージェント (CLエージェント) である。1つは、エージェントの状態が、各エージェントが存在するセルの組み合わせとして定義される。このエージェントをCL(cell)エージェントと表記する。もう1つは、エージェントの状態が、各エージェントが存在するエリアの組み合わせとして定義される。このエージェントをCL(area)エージェントと表記する。CLエージェントは、DREエージェントやRLエージェントと異なり、協調行動を達成することをタスクとし、協調行動を達成するための状態価値を学習する。CLエージェントにおいては、協調行動が成功した場合には正の報酬 R が、失敗した場合には負の報酬 D が、両方のエージェントに与えられる。衝突時には、罰 D は与えられない。つまり、DREエージェントとCLエージェントは、(1)報酬と罰の与え方、(2)状態の定義、(3)情動価値の有無の3点が異なる。CLエージェントは、 ε_e ではなく ε_s を用いて、状態価値に基づく行動決定を行う。

RLエージェントとCLエージェントの状態価値の学習には、DREエージェントと同様に、モンテカルロ法を用いた。DREエージェントとRLエージェントの

図3 エージェントエリアと情動状態

比較は、情動価値に基づく判断の影響を調べることを目的としている。DREエージェントとCLエージェントの比較は、協調行動を目的とするCLエージェントの協調行動回数に、DREエージェントの協調行動回数がどの程度近づいているのかを検証することを目的としている。

3.1.3 実験条件

エージェントは、上下右左のいずれかの方向へ1セルずつ移動することでき、エージェントの隣のセルに目標が存在する場合に目標へ到達することができる。餌探索の場合と同様に、どちらかのエージェントが目標に到達すると、その目標は存在しなくなるものとした。

各エピソードの始めに、エージェントの位置を(1)目標位置以外、(2)他のエージェントと同じ位置から開始しない、という条件下で、ランダムに決定した。エピソードの終了条件は、(1)行動制限回数に達した場合、(2)衝突が起きた場合、(3)両方のエージェントが目標に達した場合の3つである。一方のエージェントが目標に到達しても、行動制限回数までは、もう一方のエージェントは探索を続けた。

2000エピソードを1試行として、各実験において10回の試行を行った。以下の実験結果では、10試行の平均値を示す。1試行中の100エピソード毎に、協調行動回数と衝突回数の直前100エピソードの積算値を求め、その値をもとに比較を行った。

実験において、各パラメータは、 $\varepsilon_s=0.1$, $\varepsilon_e=0.1$, $R=10$, $\alpha=0.1$, $\gamma=0.95$, $\eta=0.99$, $D=-10$, $\alpha'=0.9$, $\zeta=0.95$, $\beta=0.9$ とした。情動価値の学習係数 α' は、(1)情動状態 u は状態 s よりも状態数が多い、(2)エリアにすることによって、1つのエピソードで学習される情動状態の数が少ない、(3)最も新しい情動経験を優先させるために、罰を受けたエピソードで経験しなかった情動状態の価値は減衰されるという3つの理由から、状態価値の学習係数 α よりも大きな値に設定した。

3.2 実験1：協調行動の創発

実験1では、情動価値の利用によって、DREエージェントが協調行動を創発することができるかを確認した。図4に、行動制限回数が40回の場合の協調行動回数と衝突回数を示す。DREエージェントとCL(area)エージェントの協調行動回数は、エピソードが進むにつれ、大きく増加している。RLエージェントについては、協調行動回数は50回付近で収束している。RLエージェントは協調行動を目的としていない

図4 実験1における協調行動回数と衝突回数の比較。(a) 協調行動回数。(b) 衝突回数。

ので、初期位置から近い目標を目指し、偶然の結果として、協調行動が生じている。CL(cell)エージェントについては、協調行動を目的としているので、エピソード数とともに協調行動回数がゆっくりと増加しているが、RLエージェントの協調行動回数とほとんど変わらない。CL(cell)エージェントの状態はセルの組み合わせであるので、状態数が50, 625 (225×225)と非常に多く、状態価値が十分に学習できていないためと考えられる。一方で、CL(area)エージェントの状態は、DREエージェントの情動状態と同じエリアの組み合わせである。CL(area)エージェントは協調行動を達成するための状態価値が十分学習できており、300エピソード以降では協調行動回数が100回に近い。DREエージェントは、直接的には協調行動を目的としていないので、CL(area)エージェントに比べると、協調行動回数は少ない。DREエージェントでは、100エピソード前後では片方のエージェントのみが目標に到達する場合が多く、200エピソード前後では両方のエージェントが目標に到達できない場合が多い傾向にあった。しかしながら、300エピソードまで

に状態価値と情動価値の学習が進み、300エピソード以降では70回以上の協調行動が創発された。

衝突回数に関しては、DREエージェントとCLエージェントは、RLエージェントよりも少ない。DREエージェントには衝突時に罰が与えられ、RLエージェントには罰が与えられないので、DREエージェントの衝突回数は少なくなったと考えられる。CLエージェントは、協調行動を学習しているの、別々の目標に向かう行動を取ることによって、衝突が生じる機会が少なくなったと考えられる。

以上の結果より、情動価値を利用することによって、DREエージェントが協調行動を創発することが可能であることを確認した。DREエージェントは、自身への危険を回避しつつ、タスクを達成するために、利己的な行動を取っている。第三者の視点からは、この利己的な行動が利害調整するために協調して別々の目標へ行動しているように見える。本実験の結果は、情動に基づく利己的な判断によって、社会的行動である協調行動が創発する場合があることを示唆している。

3.3 実験2：環境変化の適応

実験2では、DREエージェントが環境変化に対して適応し、協調行動を創発することができるかを確認した。環境変化として、1000エピソード後に目標を図2(a)の位置から同図(b)の位置へ変更した。

図5に、行動制限回数が40回の場合の協調行動回数と衝突回数を示す。環境変化の直後では、すべてのエージェントの協調行動回数が大きく減少している。しかしながら、DREエージェントとCLエージェントの協調行動回数は、環境変化の後でも学習初期と同様に、エピソードが進むにつれて増加している。

図6に、ある1試行におけるDREエージェントの状態価値を示す。図6(a)と(b)は、それぞれ、1000エピソード後と2000エピソード後の状態価値である。各図において、目標が存在するセルの位置をTで示している。図6から目標付近の状態価値が高いことがわかる。図7に、図6と同じ試行におけるDREエージェントの情動価値を示す。図7(a)と(b)は、それぞれ、1000エピソード後と2000エピソード後の情動価値である。各図において、目標が存在するエリアを太線で示している。図7(a)と(b)では、エージェント1とエージェント2の両方が16番目のエリアに存在する組み合わせで、情動価値が低い。この場合は、環境変化の前後で、DREエージェントは16番エリアで衝突により罰を受けているためであると考えられる。一方で、図7(a)では、エージェント1が4番エリア

(a)

(b)

図5 実験2における協調行動回数と衝突回数の比較。(a) 協調行動回数。(b) 衝突回数。

図6 実験2におけるDREエージェントの状態価値の一例。(a) 1000エピソード後。(b) 2000エピソード後。

に、エージェント2が21番エリアに存在する組み合わせで、情動価値が低い。この場合は、協調行動が成功せずに罰が与えられた状況であると考えられる。情動価値の分布は、状態価値に比べて目標位置への依存性が低いので、環境変化前に学習した情動価値の一部を環境変化後にも利用できたと考えられる。結果として、環境変化直後でも、衝突回数がほとんど増加していない。CLエージェントに関しても、DREエージェ

図7 実験2におけるDREエージェントの情動価値の一例。(a) 1000エピソード後。(b) 2000エピソード後。

ントの場合と同様に、環境変化前に学習した状態価値を再利用できたと考えられる。

RLエージェントの協調行動回数は、学習初期よりも、環境変化後は、増加が遅くなっている。これは、状態価値の目標位置への依存性が高いためであると考えられる。RLエージェントは、目標に到達するために状態価値を改めて学習し直す必要がある。RLエージェントの衝突回数は、環境変化直後に増加している。この時点では、RLエージェントの状態価値は、変化前の環境から変化後の環境へ対応するために、価値が遷移している状況であると考えられる。この状況が、衝突回数の増加につながったと考えられる。

実験1ではDREを用いたエージェントが、社会的行動である協調行動を創発することができることを確

認し、実験2ではDREを用いたエージェントの環境変化に対する高い適応性を確認することができた。扁桃体は、社会的な認知機能や社会的な意思決定に関して重要な役割を果たしていると考えられているが、情動と社会的行動の関係性については、未だに明らかになっていない部分も多い。本研究では、恐怖情動による割り込み機能を模擬したDREを用いることによって、エージェントが社会的行動を創発することを示した。提案手法は、情動と社会的行動の関係性を検証する上で、有効な手法であると考えられる。

4. おわりに

本論文では、強化学習と情動学習に基づいた意思決定法(DRE)を提案した。DREを用いたエージェントは、タスクを達成するための戦略的な判断だけでなく、危険を回避するための情動に基づいた判断を利用する。シミュレーション実験では、DREを用いたエージェントによる社会的行動(協調行動)の創発と環境変化への適応性を検証した。実験の結果、情動に基づく利己的な判断によって社会的行動が創発すること、環境変化への高い適応性を確認できた。今後は、行動決定に関するパラメータ ε_s と ε_e の社会的行動の創発への影響を調べる予定である。

謝辞

本研究は、高橋産業経済研究財団からの助成を受けて行われた。

参考文献

- [1] 松日榮信人, “生活を変えるロボットを目指して,” 日本ロボット学会誌, Vol.24, No.3, pp.292-295, 2006.
- [2] 岡田浩之, 大森隆司, “ロボカップ@ホーム：人とロボットの共存を目指して,” 人工知能学会誌 Vol.25, No.2, pp.229-236, 2010.
- [3] 重野純編, キーワードコレクション心理学, 新曜社, 1994.
- [4] E. A. Phelps and J. E. LeDoux, “Contributions of the amygdala to emotion processing: from animal models to human behavior,” *Neuron*, Vol.48, No.2, pp.175-187, 2005.
- [5] B. Seymour and R. Dolan, “Emotion, decision making, and the amygdala,” *Neuron*, Vol.58, No.5, pp.662-671, 2008.
- [6] R. Gupta, T. Kosciak, A. Bechara, and D. Tranel, “The amygdala and decision-making,” *Neuropsychologia*, Vol.49, No.4, pp.760-766, 2011.
- [7] 石黒章夫, 持田哲司, 青木猛, 内川嘉樹, “情動機構を用いた自律移動ロボットの行動調停の一手法,” 日本機械学会論文集, Vol.62, No.599, pp.234-241, 1996.
- [8] 瀧田航一郎, 荻原将文, “積極性と警戒心のバイアス

- を加味した強化学習,” 知能と情報：日本知能情報ファジィ学会誌, Vol.17, No.1, pp.103-111, 2005.
- [9] 久保田直行, 脇坂史帆, 小嶋宏幸, “情動モデルを用いたパートナーロボットに関する研究：仮想現実空間の構築と人間との相互作用,” 知能と情報：日本知能情報ファジィ学会誌, Vol.20, No.4, pp.449-460, 2008.
- [10] 高玉圭樹, マルチエージェント学習-相互作用の謎に迫る-, コロナ社, 2003.
- [11] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, 1998.
- [12] 森山甲一, 沼尾正行, “環境状況に応じて自己の報酬を操作する学習エージェントの構築,” 人工知能学会論文誌, Vol.17, No.6, pp.676-683, 2002.
- [13] 柴田克成, 上田雅英, 伊藤宏司, “強化学習による個性・社会性の発現・分化モデル,” 計測自動制御学会論文集, Vol.39, No.5, pp.494-502, 2003.
- [14] 土田昭司, 竹村和久編, 感情と行動・認知・生理-感情の社会心理学, 誠信書房, 1996.
- (2011年6月13日 受付)
(2011年12月7日 採録)
- [問い合わせ先]
〒808-0196 北九州市若松区ひびきの2-4
九州工業大学 大学院生命体工学研究科 脳情報専攻
堀尾 恵一
TEL: 093-695-6127
FAX: 093-695-6127
E-mail: horio@brain.kyutech.ac.jp

著者紹介

みさわ ひであき
三澤 秀明 [非会員]

平成17年北九州工業高等専門学校専攻科卒, 平成19年九州工業大学大学院生命体工学研究科博士前期課程修了。現在, 同研究科博士後期課程在学。ソフトコンピューティング手法の応用に関する研究に従事。IEEE, 電子情報通信学会会員。

ほり お けいいち
堀尾 恵一 [正会員]

平成8年九工大・情報工卒, 平成10年同大学院情報工学研究科博士前期課程修了。平成13年同博士後期課程修了。同年, 日本学術振興会特別研究員。平成15年九工大・大学院生命体工学研究科助手。平成19年九工大・大学院生命体工学研究科准教授。ニューラルネットワーク, ファジィシステム, およびその融合に関する研究に従事。IEEE, 電子情報通信学会, 日本知能情報ファジィ学会会員。

まつだ あつし
松田 充史 [非会員]

平成21年西日本工業大学工学部電気電子情報工学科卒業。平成23年九州工業大学大学院生命体工学研究科博士前期課程脳情報専攻修了。同年株式会社富士通ソーシャルサイエンスラボラトリ入社。現在に至る。在学中は主として強化学習手法と認知心理学に関する研究に従事。

Decision Making Based on Reinforcement Learning and Emotion Learning: Emergence of Cooperative Behavior with Selfish Judgment

by

Hideaki MISAWA, Atsushi MATSUDA and Keiichi HORIO

Abstract :

In this paper, we propose a decision making method based on reinforcement learning and emotion learning (DRE) for inducing social behaviors of agents. Emotion of animals has an important role in their social interactions. We attempt to incorporate emotion into decision making of agents. To make a social decision making, the DRE combines a decision based on intrinsic fear emotion with a strategic decision obtained by reinforcement learning. In simulation experiments, the effectiveness of the DRE is verified concerning the emergence of social behaviors and the adaptability to an environmental change through an unmoving target search problem.

Keywords : Emotion, Reinforcement Learning, Emotion Learning, Cooperative Behavior, Decision Making

Contact Address : **Keiichi HORIO**

*Department of Brain Science and Engineering, Graduate School of Life Science and Systems
Engineering, Kyushu Institute of Technology*

2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, JAPAN

TEL : 093-695-6127

FAX : 093-695-6127

E-mail : horio@brain.kyutech.ac.jp