ロボットのためらい：行動の遅れは学習効率を向上させ教えやすい印象を与える

The hesitation of a robot: a delay in its motion increases learning efficiency and impresses humans as teachable

田中 一晶[1*]　尾関 基行[1]　岡 夏樹[1]
TANAKA Kazuaki[1], OZEKI Motoyuki[1], OKA Natsuki[1]

[1] 京都工芸繊維大学 大学院工芸科学研究科
[1] Graduate School of Sclence and Technology, Kyoto Institute of Technology

Abstract: If robots learn new actions through human-robot interaction, it is important that the robots can utilize rewards as well as instructions to reduce humans' efforts. Additionally, "interval" which allows humans to give instructions and evaluations is also important. We hence focused on "delays in initiating actions" and changed them according to the progress of learning: long delays at early stages, and short at later stages. We compared the proposed varying delay with a constant delay by experiment. The result demonstrated that the varying delay improves learning efficiency significantly and impresses humans as teachable.

1　はじめに

将来、人の日常生活の場で、ロボットが人の仕事をサポートすることが予想される。そのようなロボットに予め必要な行動を全てプログラムしておくことは現実的ではない。よって、人と触れ合うロボットは人や環境とのインタラクションを通して新たな行動を獲得する能力が必要不可欠である。

ロボットが人とのインタラクションを通して学習する場合、人から与えられる行動教示のみをあてにして学習するだけでは人の負担が大きいため、報酬に基づく学習 (強化学習) を併用できることが望ましい。そのようなインタラクションでは、人が行動教示や評価教示 (報酬) を与えるための「間」が重要である。また、人が人とのインタラクションを通じて学習する場合には、学習者の能力に応じて、足場として簡単な学習課題から徐々に難しい学習課題を与えること (Scaffolding) が知られている [1] が、我々はロボットとのインタラクションにおいても人は足場を与えることを明らかにした [2]。人から与えられる報酬を利用してロボットが学習する場合、人から新しい課題 (足場) が与えられると報酬は増加し、ロボットが課題を達成するにつれて報酬は減少する。したがって、学習が進むにつれて行動教示や評価教示を与えるための「間」は不要になる。

そこで、我々は「ロボットが行動を実行するまでの時間」(以下、実行遅延) に着目し、学習の初期段階では長く、学習が進むにつれて短く実行遅延を変化させることでロボットの学習状態に応じて適切な教示が与えられると考えた [3]。つまり、決定した行動に自信が無ければ遅れて実行し、自信があれば即座に実行するわけである。この遅れは人が自信がないときに「ためらう」ことから発想を得ており、ロボットのためらいを表現する上で実行遅延は重要であると考えている。

[3] では、実行遅延が変化する条件と常に短い条件 (以下、せっかち条件)・常に長い条件 (以下、おっとり条件) の比較を、AIBO に「お手」を教えるタスクを通して行った。その結果、実行遅延を変化させる条件が最も学習効率が良く、人も教えやすいと感じることがわかった。せっかち条件は教示を与えるタイミングがつかみにくく、誤ったタイミングで教示が与えられてしまい学習効率が低下する。また、おっとり条件は誤ったタイミングで教示が与えられることはほとんど無いが、常に実行遅延が長いため、実験時間内に教示が与えられる回数が減少してしまう。さらに、学習が十分進んだ後も、教示を与える必要が無いにもかかわらず必ず間を置いて実行するため、人はイライラしてしまう。これらの条件に対し、実行遅延が変化する条件では、学習の初期段階では実行遅延が十分に長いため教えやすく、学習が進むにつれて正しい行動を即座に実行するようになるため人をイライラさせることも無い。

人ーロボットインタラクションを通してロボットが新たな行動や言葉の意味を獲得する研究は多く行われ

*連絡先：京都工芸繊維大学 大学院工芸科学研究科
〒 606-8585 京都市左京区松ヶ崎橋上町
E-mail:d8821007@edu.kit.ac.jp

てきた [4, 5, 6] が、適切な実行遅延の設定は重視されてこなかった。[4] では、PC 画面上の迷路内を 1 秒間隔で動作するエージェントを人が声で誘導するタスクを用いているが、1 秒という短い間隔では人の発話 (評価教示・行動教示) がどの行動に対して与えられたかを発話のタイミングから判断するのは難しいということがわかっている。我々は、一定の実行遅延をタスクに合わせて試行錯誤的に決めるのではなく、学習状態に応じて適切な時間に変化させることが重要であると考えている。また、[5] では、人がロボットの手を取って教える方法を用いているため、実行遅延を設定する必要は無いが、新しい行動を教える際には人が常に行動教示を与えなければならないため人の負担が大きい。これに対し、[6] では、行動教示と評価教示に基づく学習を併用して教えることが可能である学習システムを提案しており、これによって人の負担は軽減できる。

本研究では、行動教示と評価教示を併用して学習する方法を用い、実行遅延を学習状態に応じて変化させることで学習効率を向上させ、人に教えやすい印象を与えることを実験的に明らかにする。

2 学習アルゴリズムと実行遅延の決定方法

ここでは、本研究で採用した学習アルゴリズムと、実行遅延 (ロボットが行動を実行するまでの時間) の決定方法について説明する。

2.1 Q-Learning

本研究では、学習アルゴリズムとして Q-Learning[7] を採用する。Q-learning では、状態 s における行動 a_n の価値 $Q(s,a_n)$ を行動を行うたびに報酬 r に基づいて更新し、最適行動を学習する。状態 s で行動 a_n を実行した結果、状態 s' に遷移した際に、以下の更新式に従って実行した行動 a_n の Q 値を更新する。

$$\begin{aligned} Q(s,a_n) &\leftarrow Q(s,a_n) \\ &+ \alpha\{r+\gamma\max_a Q(s',a)-Q(s,a_n)\} \end{aligned} \quad (1)$$

ここで、$\max_a Q(s',a)$ は、遷移後の状態 s' で最大の Q 値を持つ行動 a の Q 値であり、将来得られる報酬の期待値を表している。α は学習率、γ は割引率であり、本研究では、それぞれ 0.1、0.5 とした。また、報酬 r については次節にて説明する。

また、行動の選択方法は Boltzmann 選択を採用し、Boltzmann 温度は 0.3 とした。

2.2 行動教示と評価教示

本研究では、ロボットは人から与えられる行動教示と評価教示を併用して学習する。

人がロボットに与えることができる評価教示は、「撫でる」という正の報酬 $+r$ と「叩く」という負の報酬 $-r$ の二種類とし、$r=1.0$ とした。これらの刺激の識別と、それに対する感情表出には、人が AIBO の頭のタッチセンサに触れた瞬間から、刺激に応じてその時点での尤もらしい表出 (喜ぶ、悲しむ) を行う随時表出システム [8] を利用する。

人から a_n を実行するように行動教示 I_n が与えられた際には、必ず行動 a_n を実行し、状態遷移後に正の報酬 $+r$ が与えられたときと同様に式 1 に従って、$Q(s,a_n)$ を更新する。強化学習に行動教示を併用する研究は [9] が挙げられる。この研究では、ポテンシャル関数によって動的に報酬の大きさを決定しているが、本研究では簡単のために定数 r を使用する。

また、人とロボットとのインタラクションでは、ロボットが望ましい行動を選択するようになると、それに応じて人が与える報酬は減少することがわかっている [2]。一般的な Q-Learning では、報酬が与えられなくなると Q 値はゼロに漸近してしまうが、この報酬の減少には、NNC(No News 規準：教示が与えられないことを肯定的な評価と捉える暗黙的な評価規準 [10]) による報酬を利用して現在の Q 値を維持することで対応できることが明らかとなっている [2]。本研究では、教示が与えられなかった際には Q 値が減少も上昇しないように、$\alpha=0$ として現在の Q 値を維持する。

2.3 実行遅延の決定方法

状態 s における行動 a_n の実行遅延 (ロボットが行動を実行するまでの時間)$T(s,a_n)$ は、Boltzmann 選択によって算出する行動 a_n の選択確率 $P(s,a_n)$ に従って、以下の式で動的に決定する。

$$T(s,a_n)=t_0+t_1/(1-e^{-ct_1\{0.5-P(s,a_n)\}}) \quad (2)$$

ここで、t_0 は最短の時間であり、t_1 は t_0 から延長する最大の時間である。これにより、行動の選択確率が高い行動ほど即座に実行し、行動の選択確率が低い行動は遅れて実行する。t_0 と t_1 は [3] と同様にそれぞれ $0.1[s]$、$3.0[s]$ とした。また、c は変化の度合を決める定数であるが、本研究では $c=0.4$ とした。

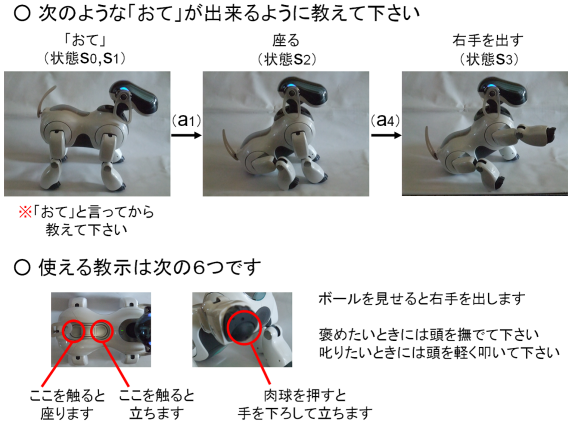

図 1: 実験参加者に提示した「おて」の手順とインストラクション

※実験参加者には () 内の文字を消したものを提示した。

3 評価実験

3.1 実験設定

[3] と同様に、AIBO が「おて」という音声を認識すると、立った状態から座って右手を出すという動作を、人から与えられる行動教示と評価教示を利用して学習するタスクを設定する (図 1)。2 節で説明した学習アルゴリズムと実行遅延の決定方法は、VC++と AIBO の開発環境の RemoteFramework を使用して実装する。

Q-Learning で使用する各状態は、以下の入力 i_0〜i_7 に従って定義する。

i_0〜i_2:それぞれ、立っている状態、座っている状態、伏せている状態であれば (1)、それ以外は (0)。

i_3〜i_6:それぞれ、左手、右手、左足、右足を出している状態であれば (1)、それ以外は (0)。

i_7:「おて」という音声を認識すると (1)、初期状態 s_0 に遷移すると (0) に初期化する。

例えば、図 1 内の 4 つの状態 (s_0〜s_3) は、初期状態では s_0 : 10000000、初期状態で「おて」を認識すると s_1 : 10000001、「おて」を認識し、座って右手を出すと s_3 : 01001001 というように定義する。以降、これらの状態は s_0〜s_3 で参照する。

また、「おて」の音声認識は AIBO の耳に実装されたマイクを介し、RemoteFramework のライブラリを利用して行う。

AIBO が各状態で実行する行動 a_0〜a_6 は、それぞれ、立つ、座る、伏せる、左手を出す、右手を出す、左足を出す、右足を出すの 7 種類とした[1]。

[1]立った状態で行動 a_0(立つ) は実行できないので、その際は状態遷移に関係しない遊び行動 (足を上げておしっこをするような振る舞いなど) を代わりに実行する。座った状態における a_1(座る) や、伏せた状態における a_2(伏せる) も同様である。また、立った状態では後足を出すことはできないので、a_5 と a_6 の代わりに足を蹴り上げる動作を行う。その際も状態遷移は起こらないものとした。

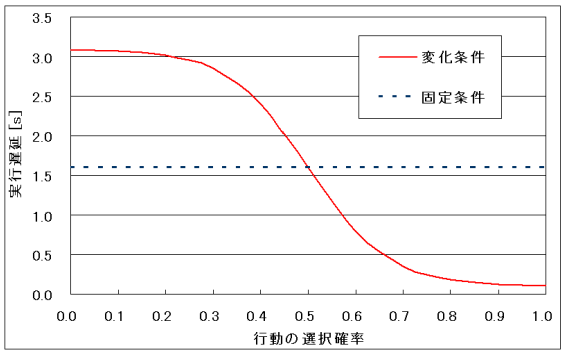

図 2: 各実験条件の行動の選択確率と実行遅延の関係

3.2 実験方法

学習状態に応じて実行遅延を変化させる方法が学習効率と人の印象に与える影響を確認するため、以下の実験条件を設定し比較実験を実施する。図 2 に各実験条件における行動の選択確率と実行遅延の関係を示す。

変化条件: T を行動の選択確率 $P(s, a_n)$ に応じ、式 (2) に従って動的に決定する。

固定条件: 式 (2) で、行動の選択確率 $P(s, a_n)$ が 0.5 のときと同じ時間 $T = 1.60[s]$ に固定する。

一つの実験条件につき 20 分間、各実験参加者で実施する順番を変えて合計 40 分の実験を実施した。実験の前には、T を 2.35[s](変化条件の最大の実行遅延 3.10[s] と固定条件の実行遅延 1.60[s] の中間) に固定し、両方の実験条件 (の最初の実行遅延) に近い条件で 10 分程度、練習を行ってもらった。

実験参加者には、予め図 1 を A4 の紙 1 枚で提示し、口頭でも紙面と同じ内容の説明を行った。実験を実施した 6 名の実験参加者の属性と実験 (変化条件・固定条件) を実施した順番を表 1 に示す。

表 1: 実験参加者の属性と実験を実施した順番

実験参加者	属性	年齢	実験の順番
A	社会人 (男性)	20 代	変化→固定
B	社会人 (女性)	50 代	変化→固定
C	社会人 (男性)	20 代	変化→固定
D	大学生 (男性)	20 代	固定→変化
E	本学学生 (男性)	20 代	固定→変化
F	本学学生 (女性)	20 代	固定→変化

3.3 ロボットの印象を評価するアンケート

実験参加者に変化条件と固定条件の印象を評価してもらうため SD 法による 5 段階評価 (1〜5) のアンケートを利用する。このアンケートは [11] で提案された、ロボットに対する人の印象を測るためのアンケートを基に作成した。[11] では、人間らしさ、生き物らしさ、好ましさ、知性、安全性の 6 つの指標と、それらを評価するための項目 (対立する形容詞) 群を示しており、

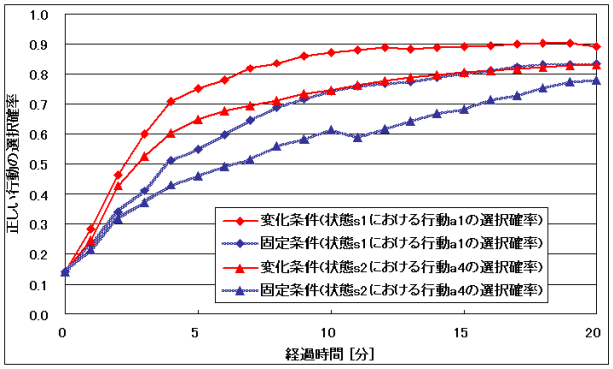

図 3: 各状態における正しい行動の選択確率の時間的な推移

本研究では生き物らしさ (6 項目)、好ましさ (5 項目)、知性 (5 項目) を評価する項目を採用した。しかし、教えやすさを評価するための項目は存在しなかったため、我々は [3] で実験参加者へのインタビューから得たロボットに対する印象を元に、教えやすさに関連しそうな 8 項目を作成し追加した。各指標に対応する項目はアンケート結果 (図 4) に示す。ただし、実験参加者にはアンケートの意図を悟られないため各項目をランダムに並べ替えたものを示した。

アンケートは 2 回の実験 (変化条件・固定条件) の後に A4 の紙 1 枚で実験参加者に提示し、先に行った実験条件は○印、後に行った実験条件は×印で同じスケール上に評価してもらった。

4 実験結果

4.1 学習結果

図 3 に状態 s_1(立っている状態) と状態 s_2(座っている状態) における正しい行動 (それぞれ a_1(座る)、a_4(右手を出す)) の選択確率の各条件ごとの時間的な推移を示す。グラフの横軸は実験の経過時間、縦軸は正しい行動の選択確率を表し、各点は実験参加者 6 名の結果を平均したものである[2]。

変化条件と固定条件の学習速度 (一定時間内に上昇した正しい行動の選択確率) の平均値の差を比較するため、実験条件 (変化条件・固定条件)、状態 (s_1・s_2)、実験の順番 (変化→固定・固定→変化) の 3 要因分散分析を行った。図 3 から分かるように、実験終了時 (20 分経過時点) では十分学習が進んでおり既に正しい行動の選択確率は収束している。そこで、分散分析での比較は学習途中である 10 分経過時点で行った。その結果、実験条件と状態に有意差があり ($F(1,4) = 14.122, MSe = 0.062, p < .05$、$F(1,4) = 85.577, MSe = 0.109, p < 0.001$)、実験の順番に差は無い ($F(1,4) = 2.152, MSe = 0.007, n.s.$) ことが示された[3]。従って、実験の順番に関わらず固定条件 (平均：0.71) よりも変化条件 (平均：0.81) の方が学習が早いことが統計的に明らかとなった。

[2]一般的に、Q-Learning での学習結果は横軸を Q 値の更新回数にして示すが、人にとってより重要なことは Q 値の更新回数を少なくすることよりも時間的に早く教えられることであると考え、横軸は実験の経過時間にして学習結果を示した。

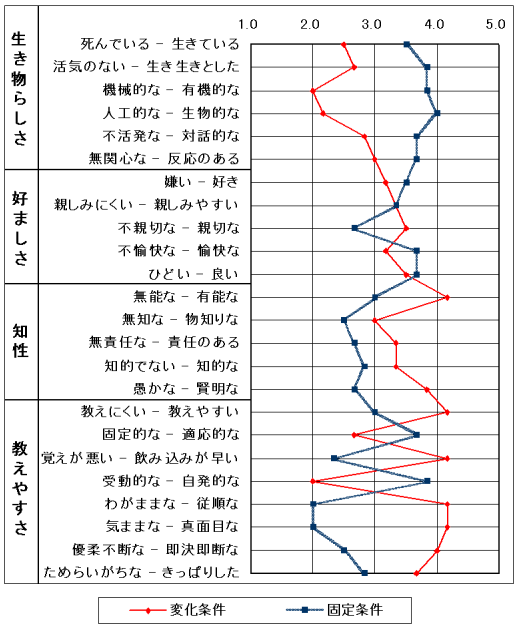

図 4: 各実験条件に対する実験参加者の印象

4.2 アンケート結果

図 4 に各実験条件 (変化条件・固定条件) に対する実験参加者の印象を示す。図上の各点は実験参加者 6 名の評価の平均値である。また、図 5 にアンケートの主成分分析結果を示す。第 1 主成分は実験条件を因子に含んでおり、変化条件と固定条件が正負の異なる因子負荷量を持つことから、第 1 主成分は変化条件と固定条件で反対の評価になった項目群を示していると考えられる (SD 法では、因子負荷量の正負を反転させると形容詞も逆になる)。一方、第 2、第 3 主成分には実験条件が因子に含まれていないため、実験条件に関わらず、相関のある項目群を示していることがわかる。

5 考察

5.1 実行遅延の変化が学習効率に与える影響

図 4 と 4 節で行った分散分析による比較から、常に実行遅延 (ロボットが行動を実行するまでの時間) が一

[3]状態 (s_1・s_2) を要因に含めたのは各状態ごとに検定を繰り返さないためであり、状態によって学習の進み方が異なるので状態にも有意差があるのは当然である。

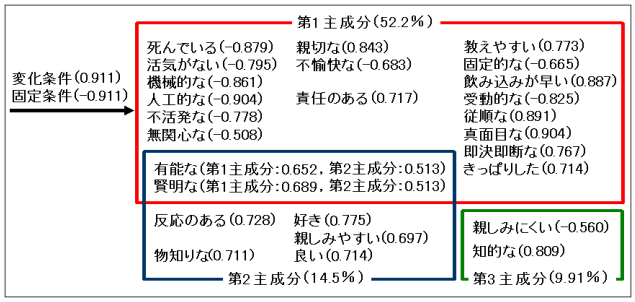

図 5: アンケートの主成分分析結果

※各主成分のラベルに付記した () 内の数値は寄与率を示し、枠内の各形容詞に付記した () 内の数値は因子負荷量を示している。本研究では SD 法を用いているため、対立する形容詞を因子負荷量の正負に応じて記載している。

定である条件 (固定条件) よりも、実行遅延をその時々の学習状態に応じて変化させた条件 (変化条件) の方が学習が早く進んでいることがわかった。これは、固定条件で設定した実行遅延 (1.6 秒) が実験参加者にとって「帯に短し、襷に長し」であったことが大きな原因である。この実行遅延は、同じ実験タスク (AIBO にお手を教える) を通して、我々の先行研究 [3] で行った実験において、短すぎて教えにくいと判断された条件 (せっかち条件) の実行遅延 (0.4 秒) と長すぎてイライラすると判断された条件 (おっとり条件) の実行遅延 (2.8 秒) の中間の時間であり、このタスクにおいて (一定の実行遅延を設定するならば) 丁度良い時間として設定した。しかし、実験参加者によって差はあるものの、学習の初期段階では誤ったタイミングで教示を与えたり、教示を与えそこねたりする状況が多かれ少なかれ起こっていた (帯に短し)。

また、学習が進むにつれて多くの実験参加者 (6 人中 4 人) が徐々に報酬 (評価教示) を与えなくなっていく。この報酬の減少は教示者がロボットに与える足場[4]の一種であり、我々の先行研究 [2, 3] でも観察されている。つまり、学習が十分に進んだ後も常に 1.6 秒遅れて行動することに意味はなく、無駄な時間になってしまう (襷に長し)。

これ対し、変化条件は学習の初期段階では十分に実行遅延が長いため (3 秒程度)、実験参加者が誤ったタイミングで教示を与えたり教示を与え損ねたりすることが少なく、学習が進むにつれて即座に行動するようになるため (0.15 秒程度)、無駄な時間も少ない。この差が変化条件と固定条件の学習効率の差となって表れたと考えられる。

しかし、変化条件において学習が十分進んだ後に探索行動 (Q 値が最も高い行動以外の行動をあえて実行する) を行う場合、実行遅延は非常に長く (3 秒程度)、無駄な時間になっていることが分かった。「何をすべきかわからない」状況ではなく「あえて探索行動を行う」場合は即座に実行すべきであり、実行した後に教示者にその評価を求めることが重要となる。つまり、あえて探索行動を行った場合には実行後に「評価を与えてもらうための時間を置く」ことが必要なのである。探索行動は即座に実行するが実行後に評価をもらうための時間を置く方法については後述の 6 節で述べる。

5.2 実行遅延の変化が人の印象に与える影響

アンケート結果 (図 4) と 4.2 節で行った主成分分析結果 (図 5) から、各条件 (変化条件・固定条件) に対する人の印象が明らかになった。実験参加者は変化条件と固定条件を相対的に評価しているため、ここでは両条件で反対の評価になった項目を示す第 1 主成分について考察する。

まず、「教えやすさ」においては 8 項目全て第 1 主成分の因子に含まれており、変化条件に対する肯定的な評価は 6 項目 (教えやすい、飲み込みが早い、従順な、真面目な、即決即断な、きっぱりした)、否定的な評価は 2 項目 (固定的な、受動的な) であった。変化条件であるにも関わらず固定的であるという評価になっているが、全ての実験参加者はその評価の理由として「覚えるのが早く、その後はほとんど同じ行動しかしない」ことを挙げていた。つまり、実行遅延の変化よりも教えた行動以外の行動 (以下、探索行動[5]) を行う頻度に着目したため、変化条件が固定的であり固定条件が適応的という評価になったわけである。変化条件が受動的であるという評価においても、固定条件が行う探索行動の頻度が比較的多いため実験参加者がこれに対して自発的だと感じ、相対的に変化条件は受動的という評価になったのであろう。つまり、これらの否定的な評価は変化条件の AIBO が早くお手を覚えた結果であり、他の 6 項目は全て変化条件に対して肯定的な評価を示しているため、固定条件よりも変化条件の方が人に教えやすい印象を与えると考えられる。「知性」に関する項目においても、同様の理由から因子に含まれる 3 項目全てが変化条件に対して肯定的な評価 (有能な、責任のある、賢明な) となったのであろう。

一方、「生き物らしさ」においては 6 項目全て第 1 主成分の因子に含まれていたが、全て変化条件に対して否定的な評価 (つまり、固定条件は全て肯定の評価) であった。これも探索行動を行う頻度が影響しており、「生物的な⇔人工的な」の項目において、固定条件を 5 または 4 と評価した実験参加者 (4 人) は変化条件よりも固

[4]足場とは、教示者が学習者の能力に応じて与える学習課題であり、簡単な学習課題から徐々に難しい学習課題を与えることを Schaffolding(足場づくり) という [1]。我々の実験で観察された報酬の減少は「報酬を与えなくてもタスクを達成せよ」という足場の一種が与えられたものと考えている。

[5]探索行動とは、Q 値が最も高い行動以外の行動をあえて実行することであり、1 − [正しい行動の選択確率] の割合で実行される。

定条件の方が様々な行動を行うことを評価の理由として挙げていた。つまり、実行遅延の変化・固定に関わらず、探索行動を多く行うことが生き物らしい印象を与えることがわかった。

「好ましさ」においては第1主成分の因子に含まれていたのは5項目中2項目のみであった。変化条件に対する好ましい印象 (親切な) は、実行遅延の変化によって教えやすかったことが起因していると考えられる。また、好ましくない印象 (不愉快な) は 20 分の実験時間中に実験参加者を退屈させてしまったことが原因と考えている。先に述べた通り、変化条件の AIBO はお手を早く覚えたため、その後はひたすら教えたお手を繰り返す作業になっていた。正しい行動の選択確率が一定以上になった時点で実験を切り上げていれば、少なくとも不愉快な印象にはならなかったであろう。

6 まとめと今後の展望

我々は、研究者がタスクに最適な実行遅延 (ロボットが行動を実行するまでの時間) を試行錯誤的に決めるのではなく、その時々のロボットの学習状態から動的に決定することが重要と考え、その方法を提案し、実行遅延が一定である条件との比較実験を通して評価を行った。その結果、我々の提案方法は実行遅延が一定である条件よりも学習効率が良く、人に教えやすい印象を与えることを明らかにした。

しかし、学習が十分に進んだ後にあえて探索行動を行う場合、長い実行遅延を置くことは無駄になるという問題も浮かび上がった。今後の取り組みとして、我々は実行遅延の決定方法を次のように改良し、その効果を確認する予定である。

実行遅延は事前遅延 T_b と事後遅延 T_a の和で決定する。

事前遅延 T_b**:** 現在の状態で取りうる全行動の選択確率から算出される平均情報量に応じて決定する。つまり、全行動の選択確率が同じ (何をすべきかわからない) 状態では平均情報量が大きいため T_b は長くなる。

事後遅延 T_a**:** 実行した行動の選択確率から算出される情報量の期待値に応じて決定する。つまり、行動の選択確率が低い行動ほど T_a は長くなる。

謝辞

本研究は科研費 (17500093 および 21500137) の助成を受けたものである。

参考文献

[1] Wood, D., Bruner, J. S. and Ross, G.: The role of tutoring in problem-solving, *Journal of Child Psychology and Psychiatry*, Vol. 17, pp. 89–100 (1976).

[2] 田中一晶, 岡夏樹：Scaffolding(足場づくり) を利用した学習系の構築, FIT2008 第 7 回情報科学技術フォーラム, RJ-006, 4 pages (2008).

[3] 田中一晶, 岡夏樹：人-ロボットインタラクションにおける「ためらう」ロボットの実験的評価, HAI シンポジウム 2008, 2B-2, 6 pages (2008).

[4] 岡夏樹, 増子雄哉, 林口円, 伊丹英樹, 川上茂雄：Fisher の直接法を用いたインタラクションデータからの意味学習, 知能と情報, Vol. 20, No. 4, pp. 461–472 (2008).

[5] Kotake, M., Katagami, D. and Nitta, K.: Acquisition of behavioral patterns depends on self-embodiment based on robot learning from multiple instructors, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, Vol. 11, No. 8, pp. 989–997 (2007).

[6] Ullerstam, M. and Mizukawa, M.: Teaching robots behavior patterns by using reinforcement learning. How to raise pet robots with a remote control, 日本機械学会ロボティクス・メカトロニクス講演会講演論文集, 2A1-L1-23 (2004).

[7] Watkins, C. J. C. H. and Dayan, P.: Q-learning, *Machine Learning*, Vol. 8, No. 3-4, pp. 279–292 (1992).

[8] 田中一晶, 岡夏樹：ペットロボットによる感情表出のタイミングがユーザとのインタラクションに与える影響, HAI シンポジウム 2006, 1B-1, 5 pages (2006).

[9] Wiewiora, E., Cottrell, G. and Elkan, C.: Principled Methods for Advising Reinforcement Learning Agents, *Proceedings of the Twentieth International Conference on the Machine Learning*, pp. 792–799 (2003).

[10] Tanaka, K., Zuo, X., Sagano, Y. and Oka, N.: Learning the meaning of action commands based on No News Is Good News Criterion, *Workshop on Multimodal Interfaces in Semantic Interaction*, pp. 9–16 (2007).

[11] Bartneck, C., Kulić, D., Croft, E. and Zoghbi, S.: Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots, *International Journal of Social Robotics*, Vol. 1, No. 1, pp. 71–81 (2009).