2B3-2

不確実性の下での人間の探索と知識利用の遷移

Transition between Exploration and Exploitation in Humans under Uncertainty

並木 尚也[*1] 大用 庫智[*1] 高橋 達二[*2]
Namiki Naoya Oyo Kuratomo Takahashi Tatsuji

[*1] 東京電機大学大学院 Graduate School of Tokyo Denki University
[*2] 東京電機大学 Tokyo Denki University

Decision making in an uncertain environment poses a conflict between the opposing demands of gathering new information and exploiting information, which is called the exploration-exploitation dilemma. It has been shown in previous studies that the loosely symmetric (LS) model is effective for handling the dilemma with the cognitive biases the model implements and correlates highly to causal intuition of humans. In this study, we compare behavioral data of humans in decision making with some representative policies in reinforcement learning to understand humans' iterative decision making under uncertainty.

1. はじめに

不確実な環境下における意思決定においては，多数の選択肢から良い選択肢を探し出す「探索」と，既知の情報・経験を活用し最良の選択肢を選択し続ける「知識利用」という 2 つの相反する行動が要求される．これを探索と知識利用のジレンマと呼ぶ．このジレンマを表現した強化学習の基本的な課題である N 本腕バンディット問題[Sutton 98]があり，この問題に対するさまざまなモデルが提案されている．その中で，人間の認知的な性質を応用して，優秀な結果を有するモデル（LS）[篠原 07]が存在する．また，脳科学の分野では人間が各選択肢を比較し相対的に評価を行っていることが明らかになっている[Daw 06]．しかしながら，人間が実際にそのジレンマに対してどのような振る舞いをするのか，あるいはどのような性質があるのかなどは具体的には明らかになっていない．

本研究では，探索と知識利用のジレンマに対して人間がどのような振る舞いをするのか，強化学習のタスクであるバンディット問題を通してさまざまな方策と比較しながら分析する．

2. 探索と知識利用のジレンマ

不確実な環境下における意思決定は，多数の選択肢から良い選択肢を探し出す「探索」と，既知の情報・経験を活用し最良の選択肢を選択し続ける「知識利用」という 2 つの相反する行動が要求される．これを探索と知識利用のジレンマと呼ぶ．得られる利益を最大にするという目的を達成するためには，このジレンマは無視できない厄介な要素である．収益を最大化するためには，最良の選択肢を見きわめ，選択し続ける必要がある（知識利用）．しかしながら，不確実な環境下では，どの選択肢が有益なのかが未知であるために一つ一つを試し検証し価値を見きわめる必要がある（探索）．知識利用を重視すると，最良の選択肢を見誤る可能性があり，結果的に目的の達成から遠ざかってしまう．探索を重視すると，利益の回収が遅れてしまい，制限のある環境下（たとえば時間，資金など），あるいはその制限が透明な環境では利益の回収が不十分になり，こちらもまた目的の達成から遠ざかってしまう．現実では無制限に試行できる環境はめったになく，さまざまな要素によって制限されるだろう．そのため，目的を達成するためには，探索と知識利用のバランスをうまく保つ必要がある．

連絡先:並木: e-mail: 10rd170[at]ms.dendai.ac.jp

探索と知識利用のジレンマは人間の経験的学習・意思決定の性質と深く関わっている．このジレンマに対して人間がどのようにうまく対処しているのかを解明することは，人間の経験的学習・意思決定の性質を理解することにつながると考えられる．また，その性質を応用することによって，人工知能やロボットなどが未知の環境において自律的に学習する事を可能にするかもしれない．そのような意味で，探索と知識利用のジレンマに対する人間の振る舞いを研究することは意義のある事であると考え，本研究を行った．

3. N 本腕バンディット問題

N 本腕バンディット問題とは，強化学習のもっとも基本的な課題の一つであり，前述した探索と知識利用のジレンマを最も単純に表現する課題である．具体例として，スロットマシンを挙げて説明する．任意の N 台のスロットマシンが存在し，それぞれに異なる当たり確率が設定されており，その当たり確率に従って報酬を返す．スロットマシンのプレイヤーは得られる報酬を最大化する事を目的とする．このときプレイヤーは各腕の当たり確率を知らず，1 度に 1 つの腕を選択する．目的を達成するために，プレイヤーは各腕の中で最良の腕を探す事（探索）と，最良と思われる腕を引き続ける事（知識利用）を要求される．このように，バンディット問題は探索と知識利用の 2 つの要素を含んでおり，単純に表現している．バンディット問題とは，このように N 個の選択肢の中から逐次的に選択し，報酬を最大化するという目的のある形態をとる問題の事である．本研究では，探索と知識利用のジレンマに対する人間の振る舞いを観測する事に都合が良いため，実験のタスクとして使用した実験では，2 個の選択肢で行った．

4. 人間の探索と知識利用のジレンマの扱い方

探索と知識利用のジレンマは，強化学習の中で中心的なトピックとして研究されてきた．近年，強化学習のタスクを通して，探索と知識利用のジレンマは脳科学でも研究され初めて来た[Daw 06]．その中でも，fMRI を用いたバンディット問題をプレイ中の参加者の脳の観測により，探索と知識利用のジレンマや学習等の人間の脳内での扱われ方が，だんだんと解明されつつある．ここで，我々は探索と知識利用のジレンマと脳科学，そして，バンディット問題と関係が深い論文を二つ紹介する．Daw et al.は 4 本腕バンディット問題をプレイ中の人間の参加者の脳活動の観測によって，探索に関連する神経基質の関係と探索と収

穫の切り替えの形式的な問題を調査した. その結果, 彼らは前頭前野腹内側部(ventral medial prefrontal cortex : vmPFC)が相対的な報酬の大きさをコード化する事と探索時に前頭極が活性化する事を示した. Daw et al. は初めて, 探索と神経基質の関係を明らかにし, 探索と知識利用のモードの間の行動戦略のスイッチングを容易にするための管理機構を映す事を可能にした. Boorman et al. は, 2 本腕バンディット問題をプレイ中の人間の参加者の脳活動の観測によって, 主に二つの脳領域の活性化と探索と知識利用のジレンマの関係を調査した. その結果, 彼らは前頭前野腹内側部が選択された腕の相対的な価値をコード化することを示した. また, 前頭極が選択されていない腕の相対的な報酬確率をコード化することを示した. 彼らは, 不確実な環境に対処可能な人間の行動の柔軟性に関して, 前頭葉における計算の重要性を示した. ただし, これらの二つのバンディット問題のタスクは非定常であった.

以上から, 不確実な環境で発生する探索と知識利用のジレンマに対処するために, 人間は絶対的評価よりも相対的な評価を行っていることが分かる. その証拠に, バンディット問題をプレイ中の人間の振る舞いが相対評価を行なう SoftMax 法で最も特徴づけられている[Daw 06]. しかし, SoftMax 法の様な評価は人間には難しいと考えられる(ランダム系列を正しく認知出来ない[Tversky 74]). また, 実際に行動としてどのように表れるかは具体的に明らかになっていない.

5. 実験

5.1 実験設定

本実験はコンピュータ上で行った. 実験参加者は東京電機大学の学生 39 名である. 参加者には 2 本腕バンディット問題に取り組み, 得られる報酬を最大化するために当たり確率の高い腕を選択するように指示された. 人間の直観性をより重視するために, どれだけ試行できるか, どの腕が今までどれだけ当たったか, あるいは外れたかなどの情報はすべて参加者には分からないようにした. 先行研究では, これらの情報が可視化されている場合が多く, それは人間の純粋な直観性とは別の傾向を生み出してしまうことが考えられる.

取り組むタスクは簡単な問題と難しい問題との 2 種類ある. 簡単な問題では 2 つの腕の当たり確率をそれぞれ(0.8, 0.2)とし, 難しい問題では 2 つの腕の当たり確率をそれぞれ(0.6 0.2)とした. 参加者の可能な試行回数は, 簡単な問題は 20 回, 難しい問題は 40 回にそれぞれ設定した. 本研究では, 最初に簡単な問題を行った後に難しい問題を行う群を ED 群と呼ぶ. 逆に最初に難しい問題を行った後に簡単な問題を行う群を DE 群と呼ぶ. 参加者をその 2 群に分けて実験を行った. また, いくつかの方策と人間を比較した.

5.2 人間と比較する方策

人間のデータと比較する方策をここで紹介する. また, スロットの評価値は客観的な条件付確率によって算出される.

(1) Greedy 法

この方策は, 選択肢それぞれの評価値に基づいて, 常に一番評価値が高い選択肢を選択する方策である. Greedy というのは貪欲という意味である.

(2) ε -greedy 法

このモデルは, 探索と知識利用の行動を明確に分離する方策である. 具体的にはパラメータ ε (0.0～1.0 の間をとる)の確率でランダムに選択肢の選択をし, 1- ε の確率で greedy に選択を行う. ε -greedy 法にはいくつか種類があり, 今回はその中の 3 つの方策を比較対象として使用した.

- 序盤探索法(Epsilon First)
  序盤探索法は, 定められた挑戦可能な試行回数の ε の割合だけ完全にランダムに選択を行う方策である.
- ε -一定法(Epsilon Constant)
  ε -一定法は, 最初の試行から最後の試行まで ε の確率が変化しない方策である.
- ε -減衰法(Epsilon Decreasing)
  ε -減衰法は, 試行回数を重ねるごとに徐々に ε の確率が減衰してゆく方策である. 本研究で用いた減衰式を以下に示す. τ は減衰のスピードのパラメータ, t はその時点までの試行回数である.

$$\varepsilon = \frac{1.0}{1.0 + \tau * t} \tag{1}$$

(3) SoftMax 法

SoftMax 法は, 条件付き確率によって算出されたスロットマシンそれぞれの評価値を選択確率に重みづけし, 選択を確率的に行うモデルである. 探索と知識利用の行動をバランシングする方策である. 本研究では, SoftMax 法を拡張した Modefied SoftMax Algorithm を使用した[Oyo 13]. 以下に式を示す. $P(X)$ はある選択肢の選択確率, $M(1|X)$はある選択肢 X に対する評価, τ は減衰率, t は現在までの試行回数である.

$$P(X) = \frac{exp(M(1|X) \times \tau t)}{\Sigma_{x' \in \{A,B\}} exp(M(1|X') \times \tau t} \tag{2}$$

5.3 実験結果

人間の探索と知識利用を観測するために,「Win-Shift」という指標を用いる. Win-Shift とは, ある腕を選択し, 当たったにも関わらず次の試行では違う腕を選択する確率である. この行動は知識利用とは最もかけ離れた行動であり, そのような意味ではある種の探索行動とみなせる(単純に腕を切り替えることも探索行動とみなせるが, Win-Shift はより探索的行動とみなせる). 図 1~4 にそれぞれの群における Win-Shift の分類, 表 1~4 に図 1~4 に対応したそれぞれのタイプとモデルごとの正解率とそのタイプの割合を示す.

Win-Shift を各個人のデータで分類した理由は, 平均化する事によりデータがつぶれ性質が見えなくなるためである. また, Win-Shift が発生したステップの期間によって分類を行っている. 正解率とは 1 回目の試行から最後の試行までの, 当たり確率の高い腕を選択した割合である.

表 1 . 簡単な問題における ED 群の正解率とタイプの割合

タイプ/モデル	正解率(%)	タイプの割合(%)
Ⅰ	74	29
Ⅱ	89	35
Ⅲ	68	11
Ⅳ	93	11
Ⅴ	80	7
Greedy 法	93	
序盤探索法	83	
ε 一定法	93	
ε 減衰法	84	
SoftMax 法	77	

表 2 . 簡単な問題における ED 群の正解率とタイプの割合

タイプ/モデル	正解率(%)	タイプの割合(%)
Ⅰ	34	11
Ⅱ	63	35
Ⅲ	82	18
Ⅳ	73	18
Ⅴ	48	6
Ⅵ	63	6
Greedy 法	72	
序盤探索法	72	
ε 一定法	69	
ε 減衰法	73	
SoftMax 法	69	

表 3 . 簡単な問題における DE 群の正解率とタイプの割合

タイプ/モデル	正解率(%)	タイプの割合(%)
Ⅰ	80	14
Ⅱ	72	14
Ⅲ	58	9
Ⅳ	58	9
Ⅴ	70	5
Ⅵ	50	5
Ⅶ	87	45

表 4 . 難しい問題における DE 群の正解率とタイプの割合

タイプ/モデル	正解率(%)	タイプの割合(%)
Ⅰ	63	14
Ⅱ	60	14
Ⅲ	88	5
Ⅳ	86	9
Ⅴ	60	23
Ⅵ	65	9
Ⅶ	85	27

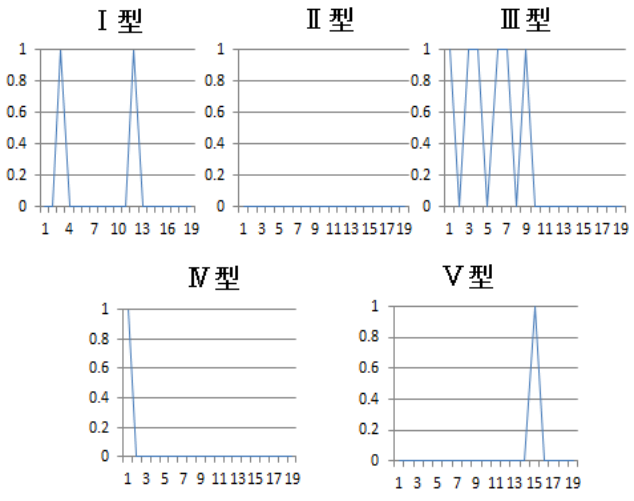

図 1. 簡単な問題における ED 群の Win-Shift の分類

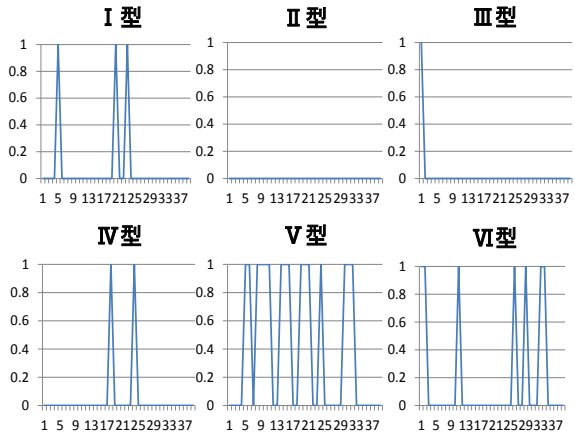

図 2. 難しい問題における ED 群の Win-Shift の分類

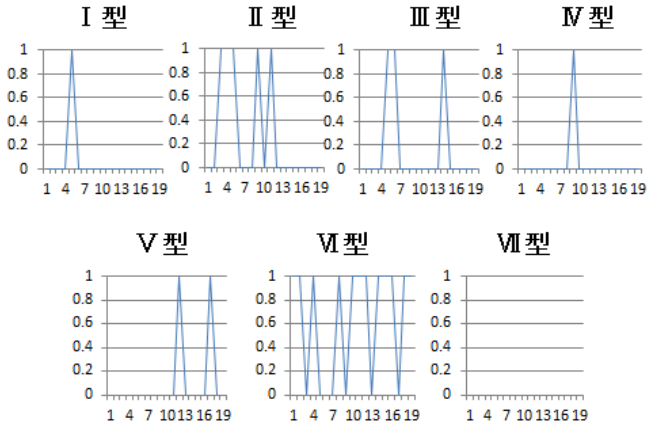

図 3. 難しい問題における DE 群の Win-Shift の分類

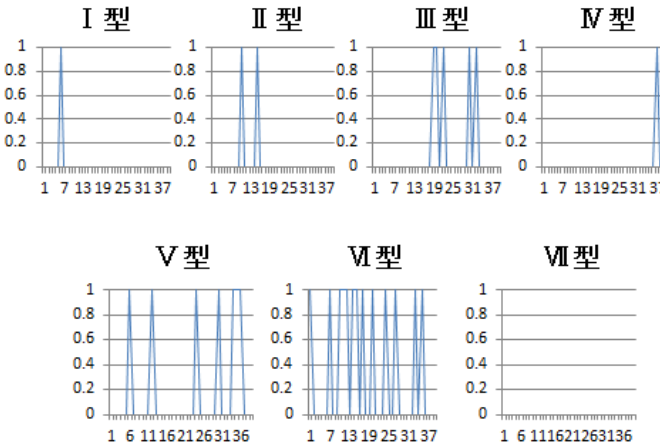

図 4. 難しい問題における DE 群の Win-Shift の分類

5.4 考察

表 1~4 より, 人間において最も多いタイプが Win-Shift が無いタイプであった. 先に紹介した greedy 法以外のモデルのように Win-Shift が出現することはなかった. したがって, 一般的に人間は良い情報, ここでは勝った後に選択肢を切り替える行動はしない可能性が高い. 逆に悪い情報, ここでは負けたという情報が選択肢の切り替え, つまり探索を促している可能性があるといえる. ただし, 単純に選択肢を切り替えることを探索と定義した場合ではあるが. また, 選択が確率的に決定しない可能性があるともいえる. 選択が確率的ならば, 勝ったあとに選択肢を切り替える行動があっても良いはずである. それにも関わらずそのような行動が一切見られないのは, やはり選択が確率的ではないといえるであろう. そもそも人間はランダム系列を正しく認識しづらく, 何らかの規則性をランダム系列に対して誤って見出してしまう[Tversky 74]. そのような性質をもった人間がランダム性をうまく扱っているとは考えにくい. 脳科学の分野で SoftMax 法と類似した傾向があるとは報告されているものの, 実際の行動データを見ると振る舞いが異なっていると考えたほうが良いと考えられる. また, 各スロットマシンに対してある程度のサンプル数を集め傾向をみてから, 知識利用の行動に移っているわけでもないように思われる. つまり, 明確に探索と知識利用の行動を分離してはいないと考えられる.

6. 結論

本研究では, 探索と知識利用のジレンマに対する人間の振る舞いの性質・傾向を調査した. その結果, 探索と知識利用の行動を明確に分ける方策や, 人間と相関があるといわれている SoftMax 法などの方策とは違う傾向があることが確認できた. 確率的に選択が行われないのである. また, 探索行動, 選択肢の切り替えは一般的には負の情報がもたらされた時のみ起こることも確認できた. これらの結果は, 現在研究されている人間の認知的な特性を利用するモデルに対して, より詳細な形式化が可能にする可能性がある.

今後の課題として, サンプル数を増やす事と, さまざまな環境設定の上で同じ結果が確認できるかどうかを検証する必要がある. 本研究では, 問題の確率の設定が片方の選択肢が高く, もう一方は低く設定されていた. そのため, 2 つの選択肢が両方とも低確率, あるいは高確率などの環境設定などでも試す必要がある. また, 今回の結果から人間の方策は Greedy 法と同傾向にあったことが確認できたが, 実際に条件付確率で評価値を算出する条件下での Greedy 法と完全に一致しているかは不明である. 一致していなければ, 人間の評価の算出の仕方が条件付き確率とは異なり, 人間特有の評価方法が存在している可能性が確認できる. 本研究では人間の方策に関して焦点を当てたが, 今後はそのような人間の評価方法に関しても調査することを課題としたい.

参考文献

[Auer 02] Auer, P., Cesa-Bianchi, N., Fischer, P., Finite-time analysis of the multi-armed bandit problem, *Machine Learning*, 47, 235-256, 2002.

[Boorman 09] Boorman, E.D., Behrens, T.E., Woolrich, M.W., Rushworth M.F., 2009. How Green Is the Grass on the Other Side? Frontopolar Cortex and the Evidence n Favor of Alternative Courses of Action. *Neuron*, 62(5), 733-743.

[Cohen 07] Cohen, J. D., McClure, S. M., Yu, A. J., 2007. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos Trans R Soc Lond B Biol Sci*, 362(1481), 933–942.

[Daw 06] Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., Dolan, R. J., 2006. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095), 876–879, 2006.

[並木 14] 並木 尚也, 高橋 達二, 探索と知識利用のトレードオフに対する人間の行動, 情報処理学会第 76 回全国大会講演論文集,517-518. (2014)

[西村 12] 西村友伸, 大用庫智, 高橋達二, 可変参照型緩対称性推論のモンテカルロ木探索での効果, ゲームプログラミングワークショップ 2012 論文集(2012‐11‐09), 2012(6), 191‐196.(2012).

[Oyo 13] Oyo,K., Takahashi,T. A cognitively inspired heuristic for two-armed bandit problems: The loosely symmetric (LS) model. *Procedia Computer Science* 24 (2013) 194-204, 2013.

[大用 11] 大用 庫智, 甲野 佑, 高橋 達二, 非定常 N 本腕バンディット問題に対する人間の認知バイアスの適用, JSAI 2011(2011 年度人工知能学会全国大会(第 25 回))予稿集, 1G1-2in, (2011).

[篠原 07] 篠原修二, 田口亮, 桂田浩一, 新田恒雄. 因果性に基づく信念形成モデルと N 本腕バンディット問題への適用, *人工知能学会論文誌*, Vol.22, No.1, pp.58-68, 2007.

[Sutton 98] Sutton, R. S., Barto, A. G., 1998. Reinforcement Learning: An Introduction. *MIT Press*, Cambridge, MA.Sidman, M. (1994). Equivalence relations and behavior: A research story. Boston, M.A.: Authors Cooperative.

[Takahashi 11a] Takahashi, T., Oyo, K., Shinohara, S., A Loosely Symmetric Model of Cognition, In: *LNCS Springer Proceedings of the 10th European Conference on Artificial Life (ECAL 2009)*, Springer, 5778, 234–241, 2011a.

[Takahashi 11b] Takahashi, T., Nakano, M., and Shinohara, S., Cognitive Symmetry: Illogical but Rational Biases, *Symmetry, Culture and Science*, 21, 1-3, 275–294, 2011b.

[Tversky 74] Tversky, A., Kahneman, D., Judgment under Uncertainty: Heuristics and Biases, Science, 185(4157), 124-1131, 1974.

[Wunderlich 09] Wunderlich, K., Rangel, A., O'Doherty, J. P., 2009. Neural computations underlying action-based decision making in the human brain. *Proc Natl Acad Sci U S A*, 106(40), 17199–17204.

[Zhang 13] Zhang, S., Yu, A.J. (2013). Cheap but Clever: Human Active Learning in a Bandit Setting. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), Proceedings of the 35th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.