2T-3

探索と知識利用のトレードオフに対する人間の振る舞い
A Behavior of Human for the Exploration-Exploitation Dilemma

並木 尚也† 大用 庫智‡ 高橋 達二‡
東京電機大学† 東京電機大学‡ 東京電機大学‡

1. はじめに

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．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）が存在する[1]．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 [2]．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モデルと比較しながら分析する．

2. 探索と知識利用のジレンマ

第１章で紹介した探索と知識利用のジレンマを具体的に説明する．不確実な環境下において，逐次的な意思決定をし，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そうそうなく，なんらかの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．また，このような実際にやってみなければ分からないという状況は人間の日常生活にあふれており，探索と知識利用のジレンマは人間の意思決定や経験的学習に関わる重要な要素である．

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，強化学習の最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．具体的には，当たり確率の異なるＮ台のスロットマシン(これを腕と呼ぶ)が存在し，プレイヤーはそのスロットマシンの中から１度に１つ選択し，決められた選択回数の中で得られる報酬を最大化することを目的とする課題である．このとき，プレイヤーは各スロットマシンの当たり確率を知らないため，実際にスロットを試し，どのスロットが有益であるかを推定する必要がある．しかしながら，得られる報酬を最大にするためには有益な選択肢を選択し続ける必要がある．

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はじめてきた．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．

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[2]．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）．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

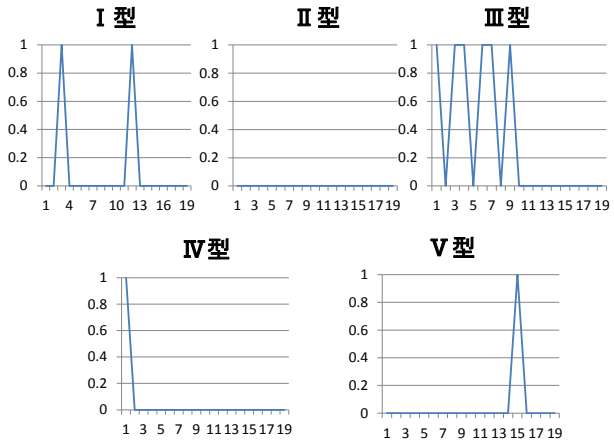

図 1. 簡単な問題における ED 群の Win-Shift の分類

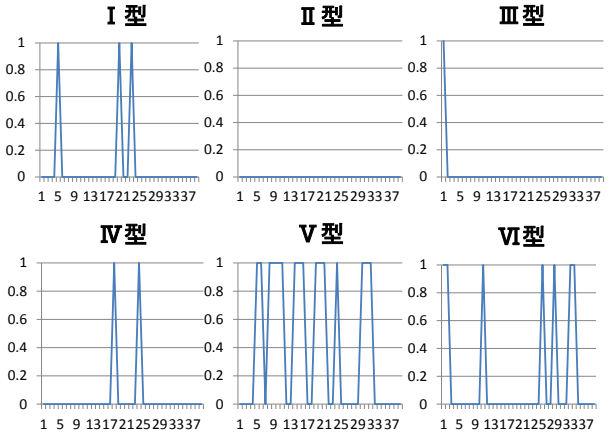

図 2. 難しい問題における ED 群の Win-Shift の分類

表 1. 簡単な問題における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
SoftMax	83	
CP	93	
LS	93	

表 2. 難しい問題における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
SoftMax	74	
CP	72	
LS	78	

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，人間と比較するために CP，SoftMax，LS の 3 つのモデルを用いる[3]．

6. 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・2 に ED 群における Win-Shift の分類，表 1・2 に図 1・2 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・2 より，簡単な問題でも難しい問題でも最も割合を占めていたタイプは Win-Shift が起きないタイプであった．Win-Shift が起きないという事は，腕の切り替えが確率的ではないという事である(腕の切り替えが確率的であれば，報酬が得られた後に違う腕を選択する可能性もある)．したがって，人間は一定の確率でランダムに探索と知識利用の行動を切り替えてはいない可能性がある(たとえば，80%の確率で知識利用行動，20%の確率で探索行動．ただし，単純に腕を切り替えることを探索行動としたとき)．つまり，探索と知識利用の行動を明確に分離せずに，混同している可能性があるといえる．

また，どちらの問題においても好成績であったのが，Win-Shift が初期の試行において見られるタイプである．特に難しい問題においてはバンディット問題において好成績を誇る LS モデルを上回る正解率を有している．このことから，初期の試行において勝ったにもかかわらず腕を切り替えるという一見非合理的な行動が良い結果をもたらし、何らかの重要な意味を持っていることが考えられる．

7. 結論

本研究では，探索と知識利用のジレンマに対して人間が一般的には探索と知識利用行動を混同している可能性がある事と，それとは別に一部分の賢い人間が無駄に思える行動により好成績を有していることが明らかになった．今後はさらに具体的な指標を用いて，上記の 2 つの性質がどのような意味を持つかを検証する必要がある．

参考文献

[1] 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.

[2] 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.

[3] 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.