1L4-OS-24b-6in

# 知識利用と探索に対する因果的直感と相対評価の処方箋的効果

## An Effect of Comparative Valuation and Causal Intuition as a Prescription for the Exploration-Exploitation Dilemma

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

[*1] 東京電機大学理工学部 1 [*2] 東京電機大学大学院
School of Science and Technology, Tokyo Denki University Graduate School of Tokyo Denki University

Decision making in an uncertain environment poses a conflict between the opposing demands of gathering and exploiting information, and that called the exploration-exploitation dilemma. It's cleared in prior research that Loosely Symmetric (LS) model is effective for the dilemma through relative intuition by simulations. We show that humans can overcome the exploration-exploitation dilemma more easily by instruction of comparative valuation as a prescription for humans.

## 1. はじめに

バンディット問題とは、当たり確率が不明の複数のスロットマシンから 1 回につき 1 つのスロットを選択し、獲得する報酬を最大にすることを目的とする強化学習のもっとも基本的な課題の一つである[sutton 98]。また、この課題は「探索と知識利用のジレンマ」(詳細は後述する)という日常的にあふれている状況を含んでおり、これは人間の意思決定に関わる重要かつ困難なものである。近年、「探索と知識利用のジレンマ」に対して人間が相対評価をしていることが脳科学で明らかになっており[e.g Daw 06]、また、シミュレーション上でも Loosely Symmetrical (LS) モデルという相対評価を含むモデルが良い成績を出している[篠原 07]。しかし人間がバンディット問題に対して相対評価を使用していることは明らかになっているが、いつどのような時に利用しているのか、あるいはどのような意図で利用しているのかは明らかになっていない。また、すべての人間が相対評価を意識して行っているわけではない可能性も考えられ、相対評価の利用がうまくできない人間が少数いることも考えられる。そのような場合に人間に対して相対評価を教示し相対評価を利用することの意識を強めたり、あるいは相対評価の利用が苦手な人間に対して相対評価を教示することによって、探索と知識利用のジレンマの克服が容易になるのではないかと考えた。

本研究では、人間が探索と知識利用のジレンマに対してどのように立ち向かうかということと、人間に簡単な相対評価を教示しそのジレンマをより克服するかどうか、という二つを目的とし、その結果を示す。

## 2. 2 本腕バンディット問題

2本腕バンディット問題の具体例としてスロットマシンを挙げて説明する。当たり確率の異なる2つのスロットマシンがあり、各々に設定された確率に従って報酬(コインなど)を出す。プレイヤーはこのスロットマシンをプレイし、報酬を最大化することを目的とする。このとき、プレイヤーはスロットマシンの当たり確率を知らない(どちらのスロットマシンが有益であるかを知らない)ため、目的(報酬を最大化する)を達成するためには有益なスロットマシンを見きわめ、選択し続ける必要がある(これを「収穫」、或いは「知識利用」と呼ぶ)。しかし、どちらのスロットマシンが有益であるかを判断するためには、いくらかどちらのスロットマシンも試行する必要がある(これを「探索」と呼ぶ)。さらにその判断をより正確にするには、より多く試行する必要がある。このように2本腕バンディッド問題は「知識利用」と「探索」という2つの重要な要素を含んでおり、この2つの要素は対立する関係にある。知識利用を重視するとどちらのスロットマシンが有益であるかを見誤る可能性があり、結果的に目的の達成には及ばなくなってしまうかもしれない。また、探索を重視すると目的を達成するための収穫をすることが遅れてしまい制限のある環境(たとえば時間、資金など)、或いはその制限が不透明な環境では十分な結果を得られなくなる可能性がある。現実では無制限に挑戦できる環境というのはそうそうなく、たいていは時間などの要素によって制限されるだろう。このような2つの要素の関係を「探索と知識利用のジレンマ」、或いは「早さと正確さのトレードオフ」と呼ぶ。2本腕バンディッド問題には、この探索と知識利用のジレンマが内在しており、プレイヤーはそれをうまく克服する必要がある。

このように2本腕バンディット問題は人間の意思決定などに関わる困難な状況の本質を含んでおり、それを再現しているものの一つだと考えられる。したがって人間の意思決定の本質・過程を観測することに都合の良い課題であると考え、今回の実験に使用した。

### 2.1 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[Boorman 09]。その中でも、fMRI を用いたバンディット問題をプレイ中の被験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 らは、4 本腕バンディット問題をプレイ中の人間の被験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 ventromedial prefrontal cortex (vmFPC)が相対的な報酬の大きさ(reward magnitude)をコード化する事と探索時に fronto polar cortex (FPC)が活性化する事を示した。Daw ら 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 らは、2 本腕バンディット問題をプレイ中の人間の被験者の脳活動の観測によって、主に

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

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 vmFPC 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,FPC 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、prefrontal computations 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ソフトマックス法で最も特徴づけられている[Daw 06]。しかし、ソフトマックスの様な評価は人間に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)。そのため、人間が何時どのような時に相対評価を上手く利用するのか、それとも相対評価以外の評価方法を混合しながら問題に適合するのかの疑問が残る。

今回の実験では、人間が探索と知識利用のジレンマに対してどのように立ち向かっているのか、また、シミュレーション上で有効な結果を出している評価方法をプレイヤーに教示することによってジレンマを克服できるのかを検証するのが目的である。

## 3. 教示する評価方法

ここでは本研究で使用する評価方法について説明する。また、以下の表は教示モデルの式に使用する表である。

**表 1：2×2 の分割表と共変動情報**

| | 試行結果 | |
|---|---|---|
| | 当たり | 外れ |
| スロット A | *a* | *b* |
| スロット B | *c* | *d* |

*a*:スロット A での当たり回数
*b*:スロット A での外れ回数
*c*:スロット B での当たり回数
*d*:スロット B での外れ回数

### 3.1 絶対評価(CP)

絶対評価とは、複数の評価対象をそれぞれ独立に評価する評価方法である。上記の2本腕バンディット問題のスロットマシンの例で説明する。2つのスロットマシン A, B があったと仮定しよう。プレイヤーは実際にこの2つのスロットマシンをプレイしてみる。仮にスロットマシン A を選択し、当たって報酬が出たとする(表 1 の a にあたる部分に 1 プラスする)。その場合、プレイヤーにとってスロットマシン A の評価は上がる。そのときにスロットマシン A の評価が上がったからといっても、スロットマシン B の評価に影響を及ぼさない(逆も然り)。つまり、実際にプレイしているスロットマシン以外は見ない・考慮しない。このようにある評価対象の評価が他の評価対象の評価に対して影響を互いに及ぼさない方法が絶対評価である。シミュレーションで用いられている式は以下のとおりである(以下の各変数は表 1 を参照)。

$$スロットマシンAの価値 = P(当たり|スロットA) = \frac{a}{a+b} \quad (1).$$

$$スロットマシンBの価値 = P(当たり|スロットB) = \frac{c}{c+d} \quad (2).$$

今回の実験では、有効性がある相対評価(RS)との比較として被験者に教示した。

### 3.2 相対評価(RS)

相対評価とは、複数の評価対象を常に比較して評価する評価方法である。上記の絶対評価の項目と同じように、2本腕バンディット問題のスロットマシンの例で説明する。同じように2つのスロットマシン A, B があったと仮定し、プレイヤーは実際にこの2つのスロットマシンをプレイしてみる。仮にスロットマシン A を選択し、当たって報酬が出たとする(この場合も表 1 の a に 1 プラスする)。その場合、プレイヤーにとってスロットマシン A の評価は上がる。そのときにスロットマシン B の評価は下がる。スロットマシン A の評価がスロットマシン B の評価に影響を及ぼす。つまり、実際にプレイしているスロットマシンについても見る・考慮する。このように、各々のスロットマシンについて1つずつ評価するのではなく、評価対象全体を見て評価し、また、ある評価対象の評価が他の対象評価に対して影響を互いに及ぼすのが相対評価である。シミュレーションで用いられている式は以下のとおりである(以下の各変数は表 1 を参照)。

$$スロットマシンAの価値 = \frac{a+d}{a+b+c+d} \quad (3).$$

$$スロットマシンBの価値 = \frac{b+c}{a+b+c+d} \quad (4).$$

## 4. 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13名である。参加者には2本腕バンディット問題に取り組ませ、当たり確率の高いスロットマシンを選択するように指示をした。その際に評価方法を教示し、それに従うように指示をした。教示する評価方法は第 3 章で挙げた相対評価(RS)と絶対評価(CP)である。教示は具体的な式を教示するのではなく、言葉とイメージ画像によって教示を行った。

取り組むタスクは 2 種類ある。1 つは、2 つのスロットマシンの当たり確率の差が大きい場合(「Big Difference」以下 BD とする)。もう 1 つは、2 つのスロットマシンの当たり確率の差が小さい場合(「Small Difference」以下 SD とする)。BD では 2 つのスロットマシンの当たり確率をそれぞれ、(0.8, 0.2)とし、SD では 2 つのスロットマシンの当たり確率をそれぞれ、(0.6, 0.4)とした。被験者の試行回数は、BD を 20 回、SD を 40 回とした。この 2 つのタスクにそれぞれの教示下で取り組んでもらう。

## 5. 結果

結果を各タスクにおける正解率と食い違い状況における教示されたモデルとの適合との 2 つの観点から見てみる。以下にそれぞれについて説明する。

### 5.1 各タスクにおける正解率

ここでは、BD・SD における各モデルの教示下における正解率を表 2 に示す。それに合わせて正解数の累積グラフを図 1 と図 2 に示す。ここでいう正解率というのは、当たり確率が高いスロットマシンを選択したかどうかの割合である。

**表 2：各状況における正解率**

| 課題 | BD | | SD | |
|---|---|---|---|---|
| 教示 | CP | RS | CP | RS |
| 正解率 | 0.75 | 0.803846 | 0.669231 | 0.665385 |

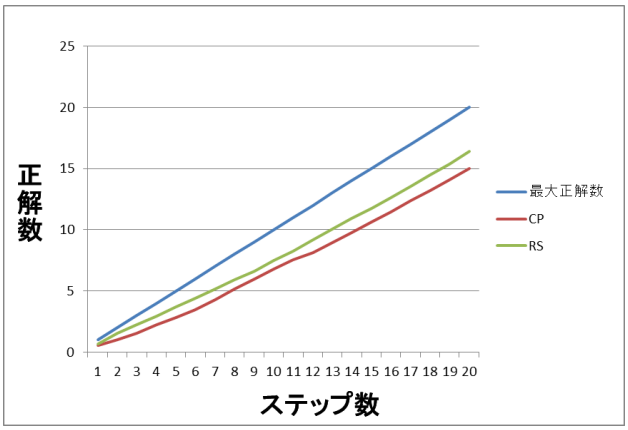


図 1：BD における正解数の累積グラフ

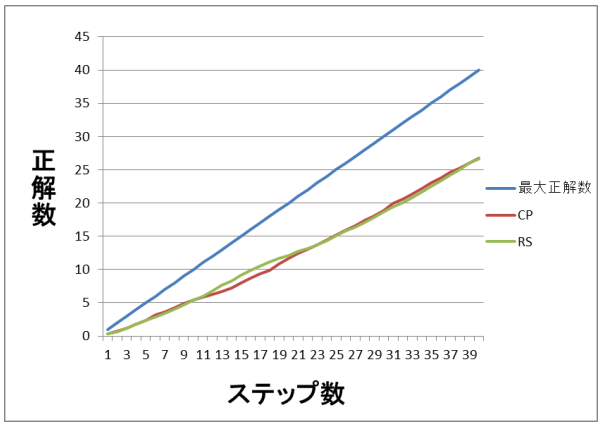


図 2：SD における正解数の累積グラフ

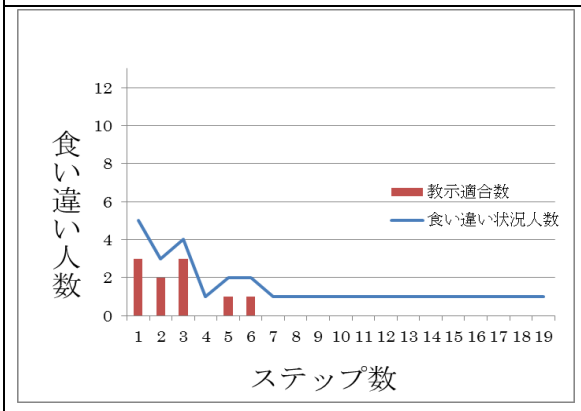


図 3：BD における CP 教示適合

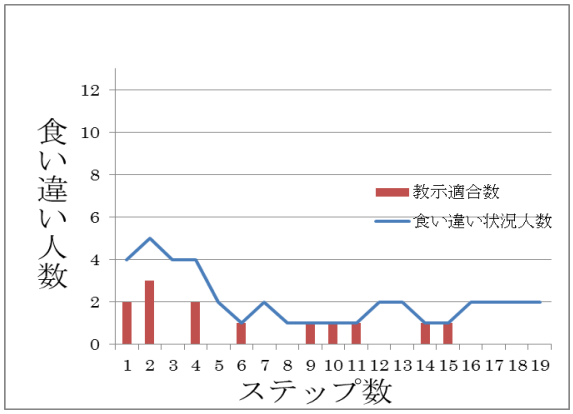


図 4：BD における RS 教示適合

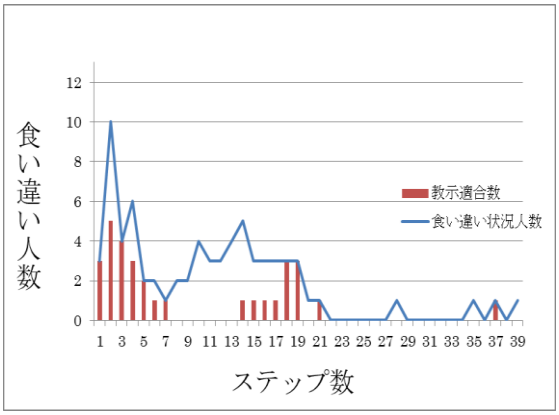


図 5：SD における CP 教示適合

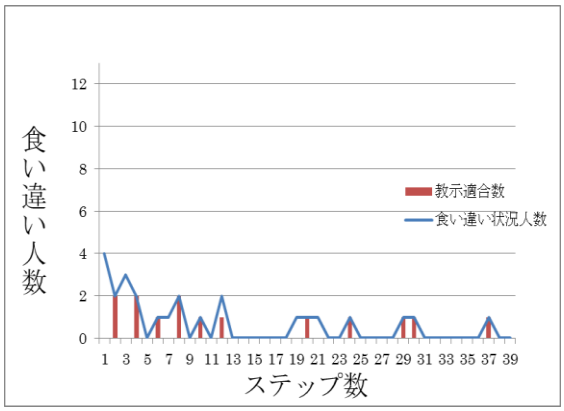


図 6：SD における RS 教示適合

表 2 から正解率が最も高いのが「課題 BD で RS 教示下」の場合である。また、図 1 と合わせて見てみると、課題 BD においては CP 教示下よりも RS 教示下の方が正解率が高く、グラフでも差がついていることが分かる。それとは対照的に表 2 と図 2 から、課題 SD において、CP 教示下でも RS 教示下でもほとんど差がない事が分かる。

### 5.2 食い違い状況における教示されたモデルとの適合

まず食い違い状況について説明する。表 1 に被験者の選択の結果を 1 ステップごとに格納し、それに基づいて 3 章の各モデルの式に代入し、各スロットの価値を更新する。このとき各モデルにおいて価値が高いスロットマシンが一致しない場合を「食い違い状況」と名付ける(図 7)。

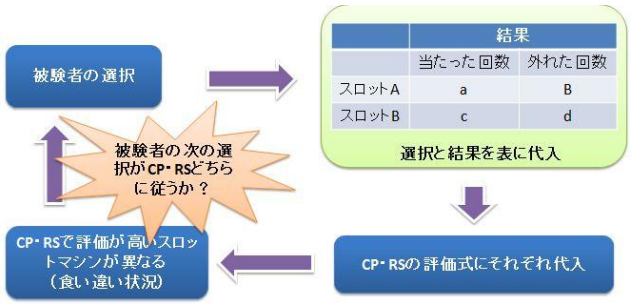


図 7 ： 食い違い状況の説明図

たとえば、3 ステップ目に CP モデルではスロットマシン A が価値が高い、RS モデルではスロットマシン B が価値が高い、という状況が食い違い状況に相当する。その食い違い状況において、ステップごとに食い違い状況に遭遇した人数と、そのとき被験者はどちらのモデルと同じ判断をするか、という事で教示効果があったかどうか、その過程を観察する。

図 3、図 4、図 5、図 6 は、図中の青い線(食い違い状況人数)が各ステップで被験者のうち何人が食い違い状況に遭遇したかという事を示している。そして、赤い線(教示適合数)がそのステップで食い違い状況に遭遇した人数のうち何人が教示に従ったかという事を示している。たとえば、図 3 であれば CP 教示下であるので、各ステップで CP モデルと同じ判断をしていれば、赤い線の値が増える。逆に図 4 では RS 教示下であるので、各ステップで RS モデルと同じ判断していれば赤い線の値が増える

図 3 と図 4 より、教示に従った人数の総数÷食い違い人数の総数は、CP が 0.33、RS が 0.325 と、全体を通してみると CP 教示下でも RS 教示下でも教示に従う人数の割合は同等である。一方、図 5 と図 6 より教示に従った人数の総数÷食い違い人数の総数は、CP が 0.43、RS が 0.56 と、RS 教示下の方が教示に従う人数の割合は若干高い。

図 3 と図 4 を見てみると、何か断定的なことは言えないように思える。一方、図 5 と図 6 を見てみると、教示に適合しているようには見えないが、面白い発見ができた。それは図 5 を見ると分かるが、7 ステップ目から 14 ステップ目において食い違い状況にあった人間全員が CP 教示に従わずに、RS と判断が一致していることが分かる。

## 6. 考察

結果より、簡単な問題(BD)において、成績が RS 教示下の方が良いように見える。しかし、食い違い状況における適合を見ると、教示の影響ではないように思える。どちらも教示適合数の割合に対して差はなかった。さらに、今回の実験では CP 教示の後に連続して RS 教示を行った。そのため、2 回目では 1 回目の経験から学習してしまい、その知識を利用した可能性が高いと思われる。

一方難しい問題においては、1 回目(CP 教示下)でも 2 回目(RS 教示下)でも成績に変化がなかった。0.2 程度の確率差はそう簡単には学習できないことが分かる。また、教示効果に関する話ではないが、新たな発見があった。タスクに取り組む過程で、あるステップ区間ですべての人間が RS 的な判断、つまり相対評価をはさんでいることが分かった。これは相対評価が難しい問題で価値を判断するために、人間にとって重要である、あるいは自然に行う行動である、という事が考えられる。また、相対評価をはさむ意味としては、探索行動から収穫行動に移るための自身の最終確認、ある種の調整、あるいはつじつま合わせのようなものだとも考えられる。つまり、探索行動と収穫行動をいきなり切り替えるのではなく、相対評価により評価対象の価値を評価する期間を保有し、行動を切り替えているのではないだろうか。

## 7. 結語

本研究は「探索と知識利用のジレンマ」に対して有効であると考えられる相対評価を人間に処方箋的に教示することによって、バンディット問題においてより良い成績を出すかどうかを調べた。その結果、教示による効果の有無ははっきりとは分からなかったが、人間の探索と知識利用のジレンマに対する選択過程・傾向を発見することができた。また、人間がどのような意味を持って相対評価を利用しているか、という事が垣間見えた。この事実は、バンディット問題に対する既存のモデルの性能の向上、あるいはまったく新しいモデルの考案へとつながり、人工知能の分野のさらなる発展へ貢献するかもしれない。これからは教示をしない場合の人間の選択過程を見ることや、効果的な教示方法の再考、そして人間が相対評価をどのような意味合いで使っているのかという部分を具体的に検証することを目標とする。

## 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.

[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.

[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, 1G1-2in, 2011.

[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 *The 17th Game Programming Workshop*. 2012.

[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.

[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.

[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.

[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 2007] 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.

[Wunderlich 2009] 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.