# ゲームAIへの認知バイアス・生物学的制約の導入に向けた人間の行動選択に関する分析

坂本 洸[1,a)]　シュエ ジュウ シュエン[1,b)]　池田 心[1,c)]

**概要**：強さを追求したゲーム AI は挙動が機械的，不自然であるなどの課題点があり，この解決のために人間らしく振る舞うゲーム AI に関する研究が行われている．「認知バイアス」や「生物学的制約」は人間らしさを再現するために導入されることがあるが，数多くあるそれらの中でどのようなものが有効であるかやその実現手法についてはまだ検証されていないものも多い．本稿では多腕バンディット問題を題材に，人間らしい振る舞いの実現に有効な認知バイアス・生物学的制約を明らかにするための分析を行う．複数のバンディット問題をプレイしてもらう被験者実験を実施し，人間の選択傾向の調査，バンディットアルゴリズムとの比較などを行った．その結果，被験者間に共通するような傾向や，バンディットアルゴリズムとは異なる挙動などが確認でき，それらに影響を与えている認知バイアスや生物学的制約について考察した．

# An Analysis of Human Behavioral Selection for Introducing Cognitive Biases and Biological Constraints into Game AI

Hikaru Sakamoto[1,a)]　Chu-Hsuan Hsueh[1,b)]　Kokolo Ikeda [1,c)]

***Abstract:*** Game AI that only focuses on strength may have some issues, such as mechanical or unnatural behaviors. To address these issues, researchers try to create game AI whose behaviors look like humans. To achieve human likeness, many approaches introduce human biological constraints and cognitive biases. Even though many approaches have been proposed, many of them have not yet been verified as to which ones are effective or how to implement them. In this paper, we target the multi-armed bandit problem to analyze which kinds of cognitive biases and biological constraints are effective in producing human-like behaviors. We conduct a subject experiment asking participants to play the multi-armed bandit problems, investigate human selection tendencies, and compare the results with other bandit algorithms. As a result, we identified common tendencies among subjects and behaviors that differ from the Bandit algorithm, and discussed the cognitive biases and biological constraints that influence these tendencies.

## 1. はじめに

現在，ビデオゲームやボードゲームの AI は人間と同等以上の強さを獲得している．深層強化学習アルゴリズムの「Agent57」では，ゲーム機 Atari2600 に搭載されている古典的ビデオゲームの 57 タイトル全てで一般的なゲームプレイヤのパフォーマンスを上回った [1]．一方，強さを追求したゲーム AI は挙動が機械的であったり，人間から見て不自然であるという課題点がある．例えば，味方キャラクタがこちらの動きに関わらず常に同じ行動をしたり，敵キャラクタが人間には到底不可能な挙動で攻撃や回避を行って人間プレイヤを倒したりした場合，人間プレイヤは違和感や理不尽さを感じゲームを楽しめなくなる可能性がある．

この課題点を解決するために，人間らしく振る舞うゲーム AI に関する研究が行われている [2]．人間らしく振る舞うゲーム AI の実現やその振る舞いの幅が広がることで，人間プレイヤから見て自然な挙動をする敵，味方キャラクタ AI や，自身と同程度の実力を持ちゲームへのモチベー

[1] 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1211，Japan
[a)] sakamotoh@jaist.ac.jp
[b)] hsuehch@jaist.ac.jp
[c)] kokolo@jaist.ac.jp

ションを高められるライバルのような存在になれる AI の実現が期待され，より人間プレイヤを楽しませることができると考える．また，得られた AI を実際の人間のような振る舞いをするテストプレイヤとして利用することで，より自然な人間の挙動を想定したデバッグが可能になることやゲーム開発のコスト削減に繋がることも期待できる．

人間らしさを実現する手法として，人間が持つ思い込みや判断の偏りである認知バイアスをゲーム AI に導入することで人間らしい振る舞いや環境変化への適応を試みる研究が行われている [3]．また，人間が生得的に持っている制約や欲求である生物学的制約を強化学習に導入する研究では，人間がゲームをプレイする際に共通して生じる現象である点で汎用性の高さに期待されている [4]．

認知バイアス・生物学的制約には視覚などの身体的なものや，記憶や感情に左右されるものなど様々な種類があるが，それらの種類によって影響を受けやすいゲームジャンルは異なったり，再現できる振る舞いは限られたりする．また，人間らしい振る舞いの創出に有効であるのか検証されていないものも多く存在する．既存の認知バイアスや生物学的制約を導入する手法では，ゲームの目標を達成する上で発生する人間らしい回避行動や失敗などを再現する試みは行われてきた．しかし，ゲームの目的に直接影響しない振る舞いに対する人間らしさや，特定の選択肢や対象へ固執するような行動などを対象としたものはあまり行われていない．

本研究では認知バイアス・生物学的制約をゲーム AI に導入し，既存手法ではできなかったような人間らしいと感じる振る舞いを獲得することを最終的な目的とする．その足掛かりとして，本稿では確率的状況における意思決定についての有名な問題である多腕バンディット問題を題材に，人間らしい振る舞いの実現に有効な認知バイアス・生物学的制約を明らかにするため人間の選択傾向について分析を行う．そして近い将来には得られた傾向を踏まえ，人間らしく振る舞っているように見えるエージェントの実装を行う．

## 2. ゲームにおける認知バイアス・生物学的制約

人間がゲームをプレイする際には，認知バイアスによって偏った判断をしたり，生物学的制約によって動作に揺らぎが発生したりする．例えば，少ない試行から得られた結果が偏っていたとしても統計的に正しいと思い込む“少数の法則”によってたまたま成功した戦略を最善だと確信してその後も取り続けることや，連続でプレイしたときの“疲れ”によって正確な操作ができなくなることなどが挙げられる．このような人間の持つ認知バイアスや生物学的制約をゲーム AI に導入する研究がいくつか行われている．

藤井らは横スクロールアクションゲームを題材に，観測時の「ゆらぎ」や認識から動作までの「遅れ」などの生物学的制約をエージェントの観測情報や報酬に対して導入した [4]．これにより，そのような人間のもつ「ゆらぎ」や「遅れ」を前提とした挙動，例えば余裕をもってゴールに向かう際の敵や障害物を安全に攻略するような人間らしい振る舞いを獲得した．だが，実際の人間のプレイにはこれ以外にも，自分が倒したい敵だけ倒して進むことや，ゴールには関係しないアイテムを収集するといった人間らしさが現れると考える．

また著者の一人は以前，より汎用的な実装を目指してゲーム固有の情報を用いずに生物学的制約を導入し，音楽ゲームのプレイヤ AI に人間らしいミスをさせることを試みた [5]．しかし，制約の種類によっては人間らしくないと評価されたものもあり，その実装手法や制約をどの程度与えるかについては改良の余地がある．

認知バイアスについては，本来人間の持つ合理的ではない認識や判断であるが，人間が限られた処理能力の中で意思決定を行うために獲得したものもあり，その性質を AI に導入することで性能の向上を図る研究も存在する．例えば，強化学習における環境の変動に脆弱であるという課題を解決するため，複数の情報の中から必要な情報のみに注意を向けるバイアスである「選択的注意」を導入する研究がいくつか行われている [3][6]．ゲームタスクに重要な部分のみに注目することにより，背景変化などに耐性を持つエージェントを実現している．しかし既存の研究では，得られる振る舞いの変化に着目したものや，人間らしい振る舞いの実現を目的としたものはまだ多くない．

エージェントに対して人間の持つ偏りを導入する研究の他に，野村らは人間プレイヤにとって自然に見える疑似乱数列を生成する際に，人間が乱数に感じる自然さの要因としてどのような認知バイアスの影響があるかについて調査を行った [7]．これも広い意味で人間らしさをゲームの改善のために用いた研究といえる．

## 3. 多腕バンディット問題

多腕バンディット問題は，複数の選択肢から 1 つを選ぶことを繰り返し，利得（報酬）の最大化を目指す問題である．スロットマシンの腕が $K$ 本存在し，各腕は選択されたときに定められた確率分布に従って報酬を返す．プレイヤは各ステップ $t = 1, 2...$ で腕を 1 つ選択して引き，得られる報酬和の最大化を目指す．このとき，報酬を最大化する上では良い腕を選択し続ける必要がある一方，良い腕がどれであるかは分からないため他の腕もある程度調べる必要があり，これを「探索と活用のトレードオフ」と呼ぶ．これはゲームプレイ時にも発生し，その際に人間は最適な行動選択を行っているとは限らず，その選択には少なからず人間らしさが現れると考える．例えば，通ったことがある道とまだ通ったことのない道のどちらへ進むか，自身が使い慣れている攻撃をするか，新しい攻撃を試してみるか

などが挙げられる．

多腕バンディット問題における人間の行動選択に関する研究として，並木らの研究があり，「報酬を得た後に腕を変更したか」という点に着目して行動選択パターンの分類を行っている [8]．また，Lefebvre らはバンディット問題をタスクとして，人間は獲得した報酬に基づいて学習する際に，予測した報酬に比べ獲得した報酬が多い場合の方が学習率が高い傾向であることを示した．そしてこれはポジティブな情報や結果に注目するポジティブティ・バイアスの影響であると考察した [9]．しかし，多腕バンディット問題における人間の行動遷移に着目したり，バンディットアルゴリズムの振る舞いをより人間らしく見えるようにすることを目的とした議論はあまり行われていない．

本稿ではバンディット問題における人間の選択にはどのような傾向や偏りが存在するのか，バンディットアルゴリズムと比較してどのような点に差異があるのかについて分析を行う．

## 4. バンディット問題を題材とした人間の行動選択に関する被験者実験

### 4.1 実験概要

多腕バンディット問題における人間プレイヤの行動選択傾向を調査するため，実際に複数パターンの問題をプレイしてもらう被験者実験を行った．ここでのパターンは，各腕の報酬確率と合計選択回数の組み合わせを指す．実験設定を以下に示す．

- 多腕バンディット問題の中でも報酬確率分布が変化しない確率的バンディット問題を対象とした．
- 実験参加者は 22〜27 歳の大学院生 10 名である．
- プレイヤは各腕を選択すると，設定した報酬確率 $p$ で“Win”(報酬 1) を，$1-p$ で“Loss”(報酬 0) を獲得する実験環境を作成した．
- 今回の実験では腕の数 $K$ は 2 本（腕 A と腕 B）として，表 1 に示す合計選択回数 ($N$) $N=50$ の試行 10 パターン，$N=10$ の試行 4 パターンについて行った．1 人の被験者につき各パターンを 2 セットずつ行い，計 28 試行実施した．このときパターンの試行順は，$N=50$ の 10 パターンをランダムに実施× 2 セット，ついで $N$=10 の 4 パターンをランダムに実施× 2 セットという順番で行った．また，セットや被験者によって腕 A と腕 B の左右を入れ替えるようなことは行えていない．
- 被験者には，獲得する報酬を最大にすることを目的として腕を選択するように指示した．また，合計選択回数は事前に告知した．
- 実験画面には「各腕を選択するボタン」と腕ごとの「直近 5 回の結果履歴」「Win の割合」「Win と Loss の回数」に加え，「現在のステップ数」「合計報酬値 (Win の合計)」を示した．このとき，Win の割合は試行後のアンケートで確率に対する質問を行うため，直近 5 回の結果履歴は直近の試行の影響について調べるため表示するようにした．実際に用いた実験画面を図 1 に示す．


**図 1** 実験画面
**Fig. 1** Experiment screen

- 被験者には各パターンの試行後にアンケートを実施し，「今回の試行について，自分の取った戦略は結果的にうまくいったと思うか」について「うまくいった / どちらとも言えない / うまくいかなかった 」の選択肢から回答してもらった．また，$N=50$，$N=10$ の 1 セット目終了時に，「どのような理由でそのような選択をしたか」「とりあえず何回か同じ腕を選ぶなど特定の挙動があったか」などを記述式で回答してもらった．今回はアンケートのへの回答を踏まえて複数の情報を実験画面に表示したが，見せるべき情報や見せ方については課題も残った．

### 4.2 比較用バンディットアルゴリズム

人間プレイヤの人間らしさを特徴づけるために，代表的なバンディットアルゴリズムを用いてその統計量を比較することにする．用いたパターンは同じ 14 種類であり，乱数性を考慮して 1 パターンにつき 10000 回の試行を行った．以下に比較するバンディットアルゴリズムを示す．

- UCB (Upper Confidence Bound)[10]
  式 (1) で求めたスコア ($\bar{\mu_i}$) が最大となる腕を選択する．$\hat{\mu_i}$ は腕 i の報酬の標本平均，$n_i$ は腕 $i$ の選択回数， $n$ は全体の選択回数である．$n_i$ が小さいほど補正項 (右辺第 2 項) が大きくなり，評価が不確かである腕を探索するようになる．また，c は探索と活用の傾向を調整する定数で， 大きいほど探索をより重視するようになる．

$$\bar{\mu_i} = \hat{\mu_i} + c\sqrt{\frac{\log n}{n_i}} \tag{1}$$

- Thompson Sampling[11]
  試行回数が有限の場合に性能が良いとされている．そ

**表 1** 実施したバンディット問題のパターン
**Table 1** Patterns of Bandit Problems Conducted

| pattern | 腕 A の報酬確率 | 腕 B の報酬確率 | step | 備考 |
|---|---|---|---|---|
| 1 | 0.25 | 0.85 | 50 | 報酬確率の差が大きい |
| 2 | 0.55 | 0.45 | 50 | 報酬確率の差が小さい |
| 3 | 0.80 | 0.70 | 50 | 報酬確率の差が小さく，確率がどちらも高い |
| 4 | 0.20 | 0.30 | 50 | 報酬確率の差が小さく，確率がどちらも低い |
| 5 | 0.65 | 0.35 | 50 | 報酬確率の差が中程度 |
| 6 | 0.05 | 0.10 | 50 | 低確率に対する過大評価の検証 |
| 7 | 0.95 | 0.90 | 50 | 高確率に対する過小評価の検証 |
| 8 | 0.50 | 0.50 | 50 | 確率が一緒である場合 |
| 9 | 0.50 | 0.50 | 50 | (特殊) 最初に 3 回連続で腕 A は当たる，腕 B は外れる |
| 10 | 0.35 | 0.55 | 50 | (特殊) 最初に 3 回連続で腕 A は当たる，腕 B は外れる |
| 11 | 0.25 | 0.85 | 10 | 報酬確率の差が大きい |
| 12 | 0.30 | 0.20 | 10 | 報酬確率の差が小さく，確率がどちらも低い |
| 13 | 0.80 | 0.70 | 10 | 報酬確率の差が小さく，確率がどちらも高い |
| 14 | 0.50 | 0.50 | 10 | 確率が一緒である場合 |

の時点での選択結果から，腕ごとに報酬期待値 $p$ の事後確率分布を求める．今回は腕が当たるか外れるかの 2 値であるためベータ分布を用いる．具体的には腕 $i$ の当たった回数を $a_i$, 外れた回数を $b_i$ としたベータ分布 $Beta(a_i+1, b_i+1)$ から乱数値を取得し，乱数値が最大であった腕を選択する．

- ε-greedy
  探索と活用のトレードオフに対する単純な解決法の 1 つである．確率 ε でランダムに腕を選択し，1- ε で現時点の平均報酬が最大の腕を選択する．

### 4.3 実験結果と分析

【一定回数連続で選択】まず，実験で得られた各被験者の行動選択の履歴を確認したところ，被験者間で共通する傾向として「序盤から中盤にかけて得た報酬にかかわらず各腕を一定回数連続で選択する」という振る舞いが多く見られた．実際に得られた行動選択履歴の一例を図 2 に示す．図の左側はパターン 1，右側はパターン 4 について，それぞれ異なる被験者の 25 ステップ目までで選択した腕と獲得した報酬を示している．「各腕を X 回ずつ選択する」という行動が複数回確認できる．特にパターン 1 の腕 A は最初に 5 回連続で外れているにも関わらず，11 ステップ目から再び 5 連続で腕 A を選択しており，報酬が得られていないから腕を変更するということよりも自身の戦略として特定の回数選択することを優先していると推測できる．

また，パターン 1 から 10 について被験者が行った計 200 試行の中から，「25 ステップまでで腕 X を 3〜10 回連続で選択後，腕 Y も 3〜10 回連続で選択かつ，腕 X の選択回数との差が± 1」という条件を満たす履歴が存在した試行を抽出したところ，59 試行確認できた．被験者ごとでそのような履歴が確認できた試行の数は 20 試行中最大で 16，最小で 1 とばらつきは大きかったものの，全ての被験者で少なくとも 1 試行はそのような行動が確認できたため，ある程度人間に共通する選択傾向の 1 つであると考える．このような傾向が見られる原因としては，人間が選択する回数に対して共通するようなチャンクを持っているのではないかと推測する．また捉え方によっては腕を毎回変更することに対して「疲れ」を感じてしまい，なるべく体力や処理能力を消費せずに意思決定を行うため，一定回数の結果を踏まえた上で選択しているのではないかとも考える．

次に，どの程度の回数連続で選択することが多いのかを調べるため，腕ごとに 3〜10 回連続して選択する回数の分布を求めた．一例として図 3 にパターン 1，3，6，7 の同じ腕を連続で引く回数の ヒストグラムを示す．横軸は各腕を連続で選択する回数で，縦軸はその合計の度数である．どのパターンにおいても，4〜5 回連続で選択する頻度が高いことが確認できる．このことから，腕の報酬確率やその時点で獲得している報酬に関わらず，一定回数連続で選択する戦略が存在すると考える．

以上のことから，各腕の選択回数や連続で同じ腕を選択する回数などに着目することにした．

表 2 に被験者の，表 3 に各バンディットアルゴリズムのパターンごとの平均報酬と正解率，腕変更回数の平均と標準偏差を示す．正解率は，各パターンの合計選択回数の中で報酬確率が高い方の腕を選択していた割合である．確率が同じパターン 8 と 9，14 については，腕 A を選択していた割合を示している．また，図 4 に報酬確率の差と被験者の正解率の関係を示す．横軸が報酬確率の差，縦軸が正解率で，図のプロットには対応するパターン番号と標準偏差を示している．以下でこれらのデータから得られた傾向について述べる．

【平均報酬と正解率】被験者と各バンディットアルゴリ

| | pattern1 | | pattern 4 | |
|---|---|---|---|---|
| step | select | reward | select | reward |
| 1 | A | 0 | A | 1 |
| 2 | A | 0 | A | 0 |
| 3 | A | 0 | B | 0 |
| 4 | A | 0 | B | 0 |
| 5 | A | 0 | A | 0 |
| 6 | B | 1 | A | 0 |
| 7 | B | 1 | A | 1 |
| 8 | B | 1 | B | 0 |
| 9 | B | 0 | B | 0 |
| 10 | B | 1 | B | 0 |
| 11 | A | 1 | A | 1 |
| 12 | A | 0 | A | 0 |
| 13 | A | 0 | A | 0 |
| 14 | A | 0 | A | 1 |
| 15 | A | 0 | A | 0 |
| 16 | B | 1 | B | 0 |
| 17 | B | 1 | B | 0 |
| 18 | B | 1 | B | 0 |
| 19 | B | 1 | B | 1 |
| 20 | B | 1 | B | 0 |
| 21 | B | 1 | A | 1 |
| 22 | B | 1 | A | 1 |
| 23 | B | 1 | A | 0 |
| 24 | B | 1 | A | 0 |
| 25 | B | 1 | A | 0 |

**図 2** 被験者の行動選択履歴の一部．別の被験者のもの (パターン 1, パターン 4)

**Fig. 2** Part of the subject's action selection history. Other subjects' (Pattern 1, Pattern 4)

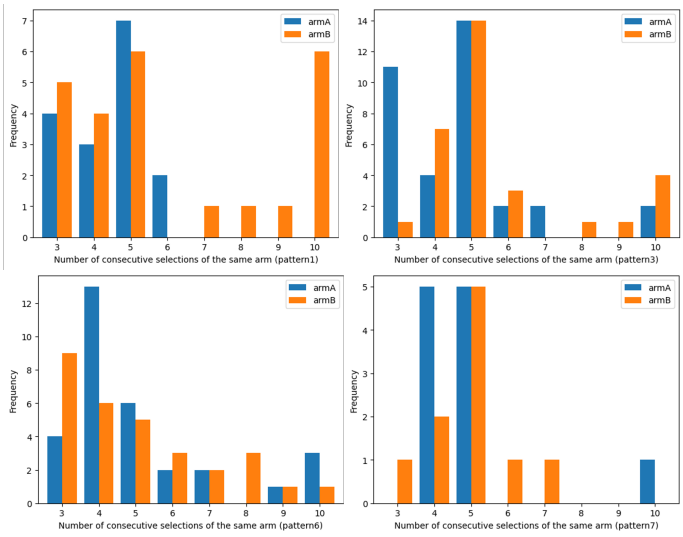


**図 3** 同じ腕を連続して引く回数のヒストグラム (パターン 1，3, 6, 7)

**Fig. 3** Histogram of the number of consecutive selections of the same arm (Patterns 1,3,6,7)

ズムの平均報酬を比較すると，人間も同程度の報酬を獲得しており，ある程度合理的な選択ができていると言える．しかし，腕の変更回数や正解率には異なる点も見られ，振る舞いには差異があると考える．図 4 から，腕同士の報酬確率差が小さいほど正解率が低い傾向にあることが分かる．また，各バンディットアルゴリズムについても同様の傾向が見られ，問題としての難易度が高くなっていると言える．パターン 2,3,4 は報酬確率は同じであるが，正解率に違いが見られた．これについては後述する．

**表 2** 被験者の平均報酬，正解率，腕変更回数の平均と標準偏差

**Table 2** Mean and standard deviation of subjects' mean rewards, correctness rate, and number of arm changes

| pattern | 平均報酬 | 正解率 | 変更回数 |
|---|---|---|---|
| 1 (0.25：0.85) | 0.75 ± 0.09 | 0.86 ± 0.11 | 6.6 ± 4.8 |
| 2 (0.55：0.45) | 0.51 ± 0.09 | 0.63 ± 0.18 | 14.4 ± 12.8 |
| 3 (0.80：0.70) | 0.77 ± 0.06 | 0.60 ± 0.30 | 8.2 ± 6.2 |
| 4 (0.20：0.30) | 0.23 ± 0.06 | 0.54 ± 0.19 | 19.6 ± 13.6 |
| 5 (0.65：0.35) | 0.60 ± 0.08 | 0.77 ± 0.16 | 10.8 ± 8.2 |
| 6 (0.05：0.10) | 0.08 ± 0.05 | 0.62 ± 0.13 | 21.0 ± 15.4 |
| 7 (0.95：0.90) | 0.93 ± 0.04 | 0.62 ± 0.36 | 8.2 ± 9.7 |
| 8 (0.50：0.50) | 0.53 ± 0.05 | (A:0.47 ± 0.26) | 12.4 ± 10.6 |
| 9 (0.50：0.50) | 0.50 ± 0.07 | (A:0.77 ± 0.17) | 10.4 ± 9.3 |
| 10 (0.35：0.55) | 0.43 ± 0.08 | (B:0.38 ± 0.22) | 13.4 ± 11.3 |
| 11 (0.25：0.85) | 0.71 ± 0.17 | 0.76 ± 0.14 | 3.3 ± 2.6 |
| 12 (0.30：0.20) | 0.23 ± 0.11 | 0.50 ± 0.13 | 6.1 ± 2.0 |
| 13 (0.80：0.70) | 0.72 ± 0.12 | 0.70 ± 0.23 | 3.7 ± 2.7 |
| 14 (0.50：0.50) | 0.46 ± 0.12 | (A:0.48 ± 0.20) | 4.6 ± 2.5 |

その中でも被験者の結果で特徴的であることとして，報酬確率の差が同じだったとしても正解率における標準偏差の差が大きいことが挙げられる．例えば報酬確率の差が 0.05 であるパターン 6 と 7 では，パターン 7 の正解率の標準偏差は 0.36 でパターン 6 の 0.13 と比べ大きく，これは ε-greedy や UCB では見られないことから人間らしい選択傾向が存在するのではないかと考えた．パターン 2，3，4 など，他の報酬確率の差が同じパターン同士を比較してみても，互いの報酬確率が高いパターンの方が標準偏差が大きい傾向にあることがわかる．

【腕の変更回数】腕の変更回数に注目すると，報酬確率の差が同じパターンの中では互いの報酬確率が高くなるほど腕の変更回数が減少する傾向となっていた．これはバンディットアルゴリズムの場合でもそうであるが，人間の場合には特に顕著であった．報酬を得た場合に腕を変更することは，得なかった場合に比べれば理由付けが難しく，報酬確率が共に高い場合に変更回数が小さくなるのはある意味自然である．しかし，人間の場合は一定以上の報酬が得られている場合にはもうその腕で満足するといった傾向が強く，このような差が生まれたと考える．

どちらの腕に満足したかどうかを調べる意味で，後半 (25 ステップ以降) に報酬確率の低い側の腕を選択し続けるという挙動が何回見られたかを調べてみた．パターン 6 では，10 人× 2 試行中，そのような挙動は一度も見られな

**表 3** 各バンディットアルゴリズムの平均報酬，正解率，腕変更回数の平均と標準偏差 (mean ± sd)

**Table 3** Mean and standard deviation of bandit algorithms' mean rewards, correctness rate, and number of arm changes

| pattern | ε-greedy ( ε =0.2 ) | | | UCB ( c=1/$\sqrt{2}$ ) | | | Thompson-Sampling | | |
|---|---|---|---|---|---|---|---|---|---|
| | 平均報酬 | 正解率 | 変更回数 | 平均報酬 | 正解率 | 変更回数 | 平均報酬 | 正解率 | 変更回数 |
| 1 | 0.75 ± 0.09 | 0.84 ± 0.13 | 9.4 ± 3.7 | 0.81 ± 0.05 | 0.93 ± 0.04 | 4.7 ± 2.2 | 0.80 ± 0.06 | 0.93 ± 0.05 | 5.5 ± 3.1 |
| 2 | 0.51 ± 0.08 | 0.61 ± 0.31 | 10.1 ± 4.0 | 0.51 ± 0.07 | 0.63 ± 0.22 | 11.5 ± 4.2 | 0.51 ± 0.07 | 0.62 ± 0.24 | 16.0 ± 6.3 |
| 3 | 0.76 ± 0.07 | 0.62 ± 0.30 | 9.9 ± 4.0 | 0.77 ± 0.06 | 0.65 ± 0.21 | 10.7 ± 5.9 | 0.77 ± 0.06 | 0.64 ± 0.26 | 14.2 ± 6.9 |
| 4 | 0.26 ± 0.07 | 0.62 ± 0.31 | 10.8 ± 4.2 | 0.26 ± 0.07 | 0.63 ± 0.17 | 17.6 ± 4.7 | 0.26 ± 0.07 | 0.62 ± 0.20 | 18.0 ± 5.5 |
| 5 | 0.58 ± 0.10 | 0.76 ± 0.23 | 9.7 ± 3.8 | 0.60 ± 0.08 | 0.82 ± 0.13 | 9.1 ± 3.5 | 0.59 ± 0.08 | 0.80 ± 0.15 | 11.6 ± 5.8 |
| 6 | 0.08 ± 0.04 | 0.60 ± 0.28 | 13.6 ± 5.6 | 0.08 ± 0.04 | 0.57 ± 0.11 | 33.0 ± 6.3 | 0.08 ± 0.04 | 0.58 ± 0.15 | 21.9 ± 4.4 |
| 7 | 0.93 ± 0.04 | 0.61 ± 0.27 | 11.4 ± 5.3 | 0.93 ± 0.04 | 0.59 ± 0.14 | 25.0 ± 12.9 | 0.93 ± 0.04 | 0.61 ± 0.28 | 14.1 ± 7.3 |
| 8 | 0.50 ± 0.07 | (A:0.50 ± 0.33) | 10.2 ± 4.1 | 0.50 ± 0.07 | (A:0.50 ± 0.24) | 11.9 ± 4.3 | 0.50 ± 0.07 | (A:0.50 ± 0.26) | 16.7 ± 6.1 |
| 9 | 0.51 ± 0.07 | (A:0.75 ± 0.22) | 10.2 ± 4.0 | 0.51 ± 0.07 | (A:0.64 ± 0.20) | 11.1 ± 3.2 | 0.51 ± 0.07 | (A:0.70 ± 0.19) | 15.2 ± 6.1 |
| 10 | 0.45 ± 0.09 | (B:0.41 ± 0.27) | 10.2 ± 4.1 | 0.48 ± 0.08 | (B:0.60 ± 0.18) | 11.6 ± 3.4 | 0.46 ± 0.08 | (B:0.50 ± 0.21) | 16.6 ± 5.3 |
| 11 | 0.67 ± 0.21 | 0.70 ± 0.30 | 2.3 ± 1.6 | 0.74 ± 0.14 | 0.82 ± 0.13 | 2.2 ± 1.6 | 0.71 ± 0.15 | 0.76 ± 0.17 | 3.2 ± 1.7 |
| 12 | 0.26 ± 0.14 | 0.55 ± 0.32 | 3.0 ± 1.9 | 0.26 ± 0.14 | 0.57 ± 0.18 | 5.3 ± 1.8 | 0.25 ± 0.14 | 0.54 ± 0.18 | 4.7 ± 1.6 |
| 13 | 0.76 ± 0.14 | 0.54 ± 0.37 | 2.2 ± 1.8 | 0.76 ± 0.13 | 0.58 ± 0.26 | 4.3 ± 2.6 | 0.76 ± 0.14 | 0.56 ± 0.27 | 3.6 ± 1.8 |
| 14 | 0.50 ± 0.16 | (A:0.50 ± 0.36) | 2.4 ± 1.7 | 0.50 ± 0.16 | (A:0.51 ± 0.24) | 3.9 ± 1.8 | 0.50 ± 0.16 | (A:0.50 ± 0.24) | 4.1 ± 1.7 |

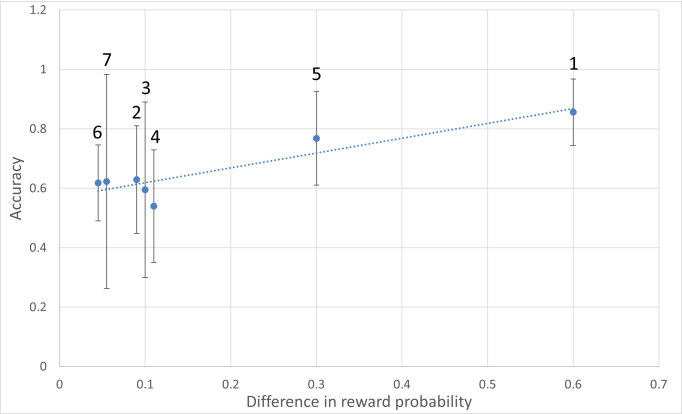


**図 4** 報酬確率の差と被験者の正解率の関係

**Fig. 4** Relationship between the difference in reward probability and the subject's percentage of correct answers.

かった．一方で，パターン 3 はパターン 6 と比べれば報酬確率の差が大きいにも関わらず，20 試行中 5 試行でそのような悪い挙動が見られた．

このような振る舞いの発生に影響を与えている認知バイアス・生物学的制約としては，自身が受け入れ可能なある程度の達成で満足するような「満足化原理」や，報酬がある程度手に入っていることから，腕を変更して報酬が得られないことを避けようとする「損失回避性」などがあると考える [12]．

【少ない結果から判断する】最初に腕 A は 3 連続で当たり，腕 B は 3 連続で外れるという特殊な処理を施したパターン 9，10 について述べる．

パターン 9 はどちらも報酬確率が 0.5 であり，通常特殊な処理をしていない場合はパターン 8 のように各腕の選択率は 0.5 程度になると考える．パターン 9 は，特殊処理の後は通常の報酬確率に戻ったものの，腕 A の選択率は 0.77 とパターン 8 よりも高くなり，腕の変更回数は減少する結果になった．この腕 A の選択率 0.77 はどのバンディットアルゴリズムの選択率よりも高かった．

パターン 10 は各腕の報酬確率が同じパターン 9 と異なり，報酬確率が 0.35 と低い腕 A が 3 連続で当たり，報酬確率が 0.55 で高い腕 B が 3 連続で外れるという特殊処理を行った．その結果，特殊処理後の報酬確率は腕 B の方が高いにもかかわらず，腕 B の選択率は 0.38 となり腕 A の選択率よりも低くなった．また，腕 B の選択率はどのバンディットアルゴリズムよりも低いという結果になった．これは，序盤に起きた偏った結果に引きずられており，少なからずその後の選択に影響を及ぼしていることが原因であると考える．また，腕の変更回数が減少したことから，序盤の結果のみでどちらの腕が良いかを判断するような傾向もあると推測する．このことから，2 章でも述べた認知バイアスである，少ない試行から得られた結果を正しいと思い込む「少数の法則」が生じていると考える．

また，パターン 10 における被験者選択履歴では，後半（25 ステップ以降）で報酬確率の低い腕 A を選択し続けるという振る舞いは 20 試行中 6 試行確認できた．加えて，腕 A を選択し続けるような振る舞いでなく，腕を交互に選択するような振る舞いであったとしても終盤まで腕 A を多く選択する傾向が見られた．これは，最初は連続で当たった腕 A が当たりにくくなったとしても，自身の持つ「腕 A が当たりやすい」という信念を修正できず固執してしまう「保守性バイアス」などに当てはまるのではないかと推測する．

表 4　パターンごとに最適な c の値
**Table 4**　Optimal c value for each pattern

| pattern | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| c | 0.05 | 0.10 | 0.10 | 0.10 | 0.00 | 0.15 | 0.00 | 0.05 | 0.40 | 0.30 | 0.70 | 1.10 | 0.05 | 1.15 |

### 4.4　人間の選択に対する UCB のパラメータ最適化

UCB について，探索と活用の傾向を調整するパラメータ $c$ がどの程度であれば人間の選択とより一致するようになるかを分析した．図 5 に $c$ を 0 から 1.5 まで 0.05 間隔で変化させた場合の，パターンごとの UCB アルゴリズムと被験者の選択一致率の平均を示す．ここでの選択一致率は，ステップごとに被験者が選択した腕と，その時点での履歴で UCB によって選択した腕が一致していたかどうかを求め，50 ステップ中で一致していたステップの割合である．また，表 4 にパターンごとに一致率が最大となった定数 c を示す．50 ステップの 10 パターンについて，一致率が最大となる c の平均は 0.125 で，どのパターンにおいても 0.5 以下となった．UCB の c としてよく用いられている $1/\sqrt{2}$ や $\sqrt{2}$ よりもかなり小さく，50 ステップの中では人間の選択は活用の傾向が強いと言える．一方，10 ステップの 4 パターンで一致率が最大となる c の平均は 0.75 でパターン 13 以外は全て 1 以上となり，ステップ数が限られた中では探索の傾向が強いという結果になった．

また，パターンごとに比較すると，パターン 1 やパターン 10 は c が小さい場合に一致率が高い傾向となった．これは腕の報酬確率の差が大きくかつ活用傾向が強い場合，最も良い腕を早めに見つけ，腕を変更しなくなる振る舞いが人間の行動選択に近いからであると考える．

またパターン 7 やパターン 3 は，c が大きくなった場合一致率が大きく減少している．この 2 つのパターンは報酬確率の差が小さくかつ互いの報酬確率が高いという共通点がある．報酬確率が互いに低い場合，どちらの腕も悪い腕だと思い，人間もある程度探索傾向にあった．一方，報酬確率が互いに高い場合，どちらの腕でも一定の報酬を得ることができるため，人間は早い段階で探索を打ち切り，一方の腕を引き続けてしまう傾向が見られた．このことが，c が大きくなると一致率が減少する原因であると推測する．しかし，現在の一致率の求め方では試行の後半で UCB の選択と不一致になり続ける場合があり，人間らしいバンディットエージェントを作るのであれば，「両方の腕を一定回数ずつ引く場合がある」「選択率が高ければ後半は妥協して同じ腕を引き続ける場合がある」といった内容を含めることが有益かもしれない．

## 5.　おわりに

本稿では，人間らしい振る舞いの実現に有効な認知バイアス・生物学的制約を明らかにすることを目標に，多腕バンディット問題を題材に人間の選択傾向について分析を行った．観測できた振る舞いや傾向について，いくつかの認知バイアスや生物学的制約の影響があることを確認した．今後は引き続き分析を進め，傾向や振る舞いを説明できる認知バイアスや生物学的制約について考察し，それらの影響を考慮した上で人間らしく見えるように振る舞う手法を提案する．そして，多腕バンディット問題に限らずさまざまなゲームに対して，人間らしい行動選択を実現する認知バイアス・生物学的制約を再現する方法論の提案・実証を行いたい．

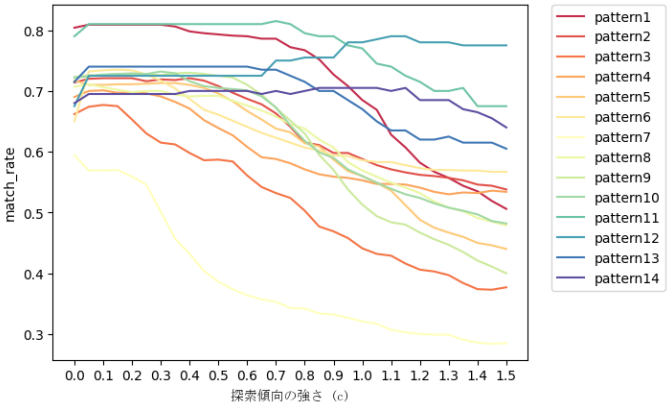


図 5　探索傾向の強さ (c) と選択一致率の関係
**Fig. 5**　Relationship between the strength of the exploration propensity (c) and the selection agreement rate

## 謝辞

本研究は JSPS 科研費 JP18H03347，JP20K12121 の助成を受けたものです．

## 参考文献

[1] BADIA, Adrià Puigdomènech, et al. Agent57: Outperforming the atari human benchmark. In: International Conference on Machine Learning. PMLR, 2020, pp. 507-517.
[2] 佐藤直之, et al. Influence Map を用いた経路探索による人間らしい弾避けのシューティングゲーム AI プレイヤ, GPW2016 論文集，pp.57-64，2016.
[3] Tang, Yujin, et al. “Neuroevolution of self-interpretable agents.” Proceedings of the 2020 Genetic and Evolutionary Computation Conference. p.414-424, 2020.
[4] 藤井叙人，et al. 生物学的制約の導入によるビデオゲームエージェントの 「人間らしい」 振舞いの自動獲得．情報処理学会論文誌 55.7，pp.1655-1664, 2014.
[5] 坂本洸，橋本剛．音楽ゲームのプレイヤ AI における人間らしく振る舞う強化学習手法の提案. 研究報告ゲーム情報

学 (GI)，2021，pp.1-7.
[6] 岩瀬諒，鶴岡慶雅．選択的注意機構を用いたロバストな強化学習手法の実現．GPW2021 論文集，pp.71-77, 2021.
[7] 野村久光，et al．標準的なゲームプレイヤにとって自然に見える疑似乱数列の生成法，GPW2013 論文集，pp.27-44, 2013.
[8] 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. 第 76 回全国大会講演論文集, pp.517-518, 2014.
[9] Lefebvre, Germain, et al.“Behavioural and neural characterization of optimistic reinforcement learning.” Nature Human Behaviour 1.4，pp.1-9，2017.
[10] Auer，Peter，at al.“Finite-time analysis of the multiarmed bandit problem.” Machine learning 47.2，pp.235-256, 2002.
[11] Agrawal, Shipra, and Navin Goyal.“Analysis of thompson sampling for the multi-armed bandit problem.” Conference on learning theory. JMLR Workshop and Conference Proceedings, pp. 39.1-39.26, 2012.
[12] 箱田裕司，et al. 認知心理学，有斐閣，2010.

 