プレイヤーの良い行動に報い，悪い行動をとがめるランダム要素調整

WANG YICHEN[1,a)]　鈴木 克磨[1,b)]　HSUEH CHU-HSUAN[1,c)]　池田 心[1,d)]

概要：ランダム要素を含むゲームでは，良い行動が不運によって報われず，悪い行動が幸運によって報われることがある．そこで本研究では，ランダム要素と戦略要素を含む一人ゲームを対象とし，プレイヤーの行動の良し悪しに応じてランダム要素を調整する手法を設計する．予備実験として，ブラックジャックと 3 × 3 盤面の 2048 において，調整した環境で強化学習エージェントを学習させ，調整のない元のゲームで評価した．その結果，両ゲームとも弱い調整でも学習は促進され，中程度の調整は良い行動の学習をより大きく促しうることが示された．一方，強すぎる調整は，一部の状態が訪問されなくなることや，元のゲームでは通用しない戦略が学習されることなどの問題を招いた．

Randomness Adjustment That Rewards Good Player Actions but Penalizes Bad Player Actions

WANG YICHEN[1,a)]　SUZUKI KATSUMA[1,b)]　HSUEH CHU-HSUAN[1,c)]　IKEDA KOKOLO[1,d)]

Abstract: In games with randomness, good actions may go unrewarded because of bad luck, and bad actions may be rewarded by good luck. In this study, we design a method for single-player games with both random and strategic elements that adjusts the randomness according to the quality of the player's actions. In preliminary experiments, we trained reinforcement learning agents in adjusted environments of Blackjack and of 2048 on a 3×3 board, and evaluated them in the original games. The results show that even weak adjustment promoted learning in both games, and that moderate adjustment can promote the learning of good actions more strongly. However, overly strong adjustment caused problems such as some states no longer being visited and the agent learning strategies that do not work in the original game.

1. はじめに

多くのゲームには，トランプの配布やサイコロの出目などに代表されるようなランダム要素が含まれている．これにより，同じプレイヤーが同じ戦略を取っても異なる展開や結果が生じ，ゲームに多様性がもたらされる．

ランダム要素はゲーム展開に変化や緊張感を与える．一方で，プレイヤーのスキルや戦略の良し悪しとは関係なく，運だけでゲームの結果が決まる場合もある．その結果，スキルの高いプレイヤーが良い行動を取ったにもかかわらず，不運によって敗北し，不満を抱くことがある（図 1）．

[1] 北陸先端科学技術大学院大学
Japan Advanced Institute of Science and Technology
[a)] s2550001@jaist.ac.jp　[b)] katsuma@jaist.ac.jp
[c)] hsuehch@jaist.ac.jp　[d)] kokolo@jaist.ac.jp

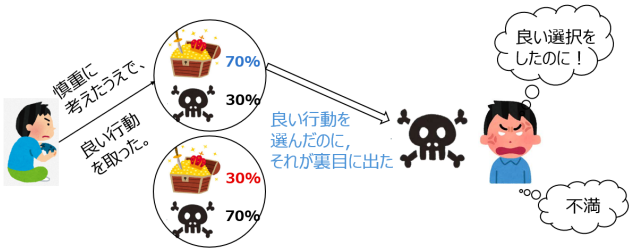


図 1　課題 1: 良い戦略が不運により報われない状況

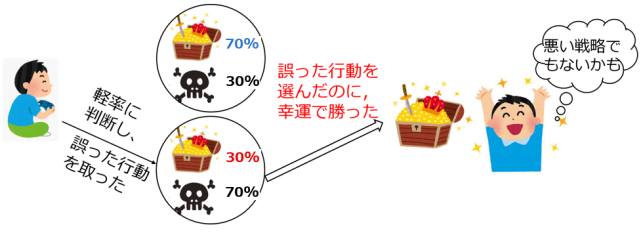


図 2　課題 2: 悪い戦略が幸運により強化される状況

© 2026 Information Processing Society of Japan

反対に，スキルの低いプレイヤーが偶然の幸運によって勝利し，悪い行動を正しいと学習してしまうこともある（図2）．これら2つの課題は，運の影響によって，ゲームの結果が行動の良し悪しを反映しなくなることから生じる．

実際のゲームでも，ランダム性が一般にもたらすこうした副作用に対処するための方法がいくつか取り入れられている．ここでは代表的な3つの方法を紹介する．

- 確率の提示: 攻撃の命中率やアイテムのドロップ率など，結果が生じる確率を数値で明示する方法である．法令などに基づいて表示されることもあり，プレイヤーはその確率を信頼して，不運な結果も起こりうるものとして納得しやすくなる．一方で，確率そのものは変えないため，個々のプレイヤーが経験する結果は保証されず，命中率90%の攻撃が続けて外れるといった極端な結果も起こりうる．
- 動的難易度調整（DDA）: プレイヤーの状況に応じて難易度を動的に調整する方法で，一般的には技量による有利・不利を緩和して没入を保つために用いられ，例えばゲーム中のアイテム供給量を調整する手法などもある [1]．運不運の補償にも使えるが，DDAは不運と悪い行動を区別しないため，悪い行動をとるほど環境が易しく，良い行動をとるほど難しくなりうる．これは本研究の目指すものとは逆の調整である．
- 天井（Pity System）: ガチャなどで望ましい結果が一定回数続けて得られなかった場合，次の試行でその結果を確実に得られるようにする仕組みである．確率の提示とは異なり，一定回数以内に望ましい結果が得られることを個々のプレイヤーに保証し，極端な不運を防ぐ．ただし，試行回数に基づく保証であり，良い行動に報いる用途には直接使いにくい．

そこで本研究では，ランダム要素と戦略要素を含む一人ゲームを対象とし，プレイヤーの行動の良し悪しに応じてランダム要素を調整するメカニズムを設計する．これにより，ゲームのランダム事象に行動の良し悪しがより強く反映されるようになることを目指す．具体的には，良い行動を取った場合には有利な結果が得られる確率を高め，納得のいく勝利による達成感と行動の強化につなげる．一方で，悪い行動を取った場合には偶然の幸運による勝利を起こりにくくして，その行動の過大評価を防ぐ．これには，敗北に納得しやすくなることで，プレイヤーが自分の行動を見直すきっかけとなる効果も期待できる．

ランダム要素の調整にはリスクも伴う．本手法は本来の確率分布とは異なる形でゲーム内のランダム事象を制御するため，プレイヤーのランダム性に対する認識を歪める可能性がある．例えば，プレイヤーが調整後の経験に基づいて，元のゲームにおける勝ちやすさや行動の価値を誤って見積もるおそれがある．このようなリスクの検証も，本研究の重要な課題の一つである．

提案するランダム要素調整メカニズムの有効性については，最終的に被験者実験を通じて評価する予定である．具体的には，ランダム要素を調整していない場合と比べて，提案手法がプレイヤー体験の観点からより望ましいかを確認したい．また，ある程度長期間のプレイを要するが，悪い行動の学習が抑えられるかも確認したい．本稿ではその前段階として，ブラックジャックや3×3の盤面を用いた2048など，構造が比較的単純なゲームを対象にして，エージェントを用いた予備実験を行う．これにより，提案手法が良い行動に報い，悪い行動をとがめるという設計意図に沿って機能するか，またその調整がエージェントの行動学習に良い影響を与えるかを検証する．

2. 関連研究

本章では，プレイヤー体験の改善と良い行動の学習支援に関する研究を整理する．ランダム性の副作用への対処法，行動の結果や評価による学習，強化学習における教示を整理し，本研究との違いを示す．

2.1 ランダム性の副作用への既存の対処法

表1に既存手法と本研究を目的，確率の改変，明示性，行動の良し悪しとの関係から比較する．

野村ら [2] は，数学的に優れた疑似乱数が一般のプレイヤーには自然に見えるとは限らないことに着目し，人間が想定するランダムな系列の偏り（同じ目の連続が少ないなど）に沿うように乱数列を最適化して，より自然に見える乱数列を生成した．また，乱数列単体とゲーム内で用いる場合とでは自然さの判断に影響する要因が異なると考え，すごろくで罠に落ちる頻度や出目について，プレイヤーが自然と感じる傾向を調べた．こうした調査に基づき，罠に落ちる頻度や連続を調整する手法を提案し，予備的な被験者実験で不満が小さくなる傾向を示した．本研究も結果そのものを調整する点で共通する．ただし，表1に示すように，既存の方法では調整がプレイヤーの行動の良し悪しと関係しないか，DDAのように逆向きになりうるのに対し，本研究は行動の良し悪しを基準に結果を調整する．

2.2 行動の結果によるフィードバック

本研究の目的の一つは，良い行動が報われやすいようにすることでプレイヤーの満足度を高めることであるが，行動の良し悪しに応じた結果を与えることは，良い行動の学習にも好影響を与えると期待する．本節では学習の観点から，プレイヤーがゲームの結果などのフィードバックを通じて行動の良し悪しを学ぶことに関する研究を述べる．フィードバックにはヒントや解説など多様な形があるが，ここではゲーム内で生じる結果と，行動の良し悪しの評価に限る．Kumariら [3] は，カジュアルゲームのプレイヤー13名へのインタビュー調査から，運の影響が大きくない場合は，プレイヤーは結果を通じて自分の技能を把握し，

© 2026 Information Processing Society of Japan

表 1 ランダム性の副作用に対処する方法の比較

方法	主な目的	確率の改変	明示性	行動の良し悪しとの関係
確率の提示	結果の起こりやすさを伝え，不運な結果への納得を得る．	しない	明示	関係しない．
DDA	運や技量の差によって難易度が上下した場合，難易度を適切に保ち，没入状態を維持する．	場合による	主に非明示	良い行動には結果的に不利な状況が与えられることが多い．
天井	極端な不運を防ぐ．	天井に達した試行のみ	明示	関係しない．
野村ら [2]	乱数を自然に見せ，操作されているという不信感を防ぐ．	乱数列や結果を調整	非明示	関係しない．
本研究	ランダムに決まるゲームの結果に行動の良し悪しを反映させ，不運による不満を防ぎ，良い行動の学習を助ける．	行動の分類に応じて調整	非明示	良い行動ほど有利になる．

誤りに気づいて改善できることを報告している．裏を返せば，運の影響が大きい場合は，結果は行動の良し悪しを知る手がかりになりにくい．

フィードバックが学習に与える効果については，Gupta ら [4] がハノイの塔において最適な価値関数に基づいて各手の良し悪しを評価し，その評価を被験者に提示すると学習が促進され，評価が提示されない課題への転移も向上することを示した．ただし，ハノイの塔はランダム要素を含まない課題で，評価も結果とは別に提示されていた．一方，本研究ではランダム要素によって，結果が行動の良し悪しを反映しにくいゲームを対象とする．評価を別途提示せず，ランダム要素を調整して結果に行動の良し悪しを反映させることで，結果自体がフィードバックとして機能するようにし，プレイヤーの学習を助けることを目指す．

2.3 強化学習における教示

本稿では，人間のプレイヤーに先立ち，経験した結果のみから学習する強化学習エージェントを用いて調整の効果を検証する．本研究の調整は学習者の経験を外部から操作するものであり，強化学習における教示の一種と捉えられる．Zhang ら [5] は，教師が学習者に目標の戦略を教える方法を，正しい行動を直接示す示範による教示（模倣学習はこれに当たる）と学習者の経験を操作する教示に分けた．後者は教師が操作できる対象によって，さらに区別され，報酬のみを操作できる場合は報酬整形に相当し，より強い場合には起こりうる遷移先の中から次の状態も選べる．Zhang らは各方法について，Q 学習が目標戦略を学ぶまでに最悪の場合に要する経験量を解析し，環境の遷移に沿って導く場合にはランダム探索のため到達しにくい状態の教示に膨大な経験が必要で，報酬のみの操作では稀な遷移先で特に顕著になることを示した．

本研究も学習者の経験を操作する教示に属するが，報酬は変えず，遷移先も直接選ばずに，出現確率だけを偏らせる．また，Zhang らが学習者の Q 値を参照するのに対し，本研究は内部状態を参照せず，元のゲームの行動価値に基づく規則に従う．その意味で本研究の調整はこれらの教示よりも間接的である．学習者は，元のゲームと同様に自ら選んだ行動の後に生じた結果からその行動の良し悪しを推し量り，調整は結果に行動の良し悪しを反映させやすくするにすぎない．そのため，調整を最大にしても正しい行動が直接示されることはなく，この点で模倣学習とは異なる．また，結果は学習者が実際に訪れた状態でしか得られないため，調整後の環境で到達できない状態は学習できないと予想される．以上の比較を表 2 にまとめる．

表 2 強化学習における教示と本研究の調整の比較

強化学習における教示	操作の内容
示範（模倣学習）	各状態で取るべき行動を直接示す．
報酬による教示（報酬整形）	学習者の Q 値に応じて，各ステップの報酬の値を設定する．
報酬と遷移先による教示	報酬の値を設定し，起こりうる遷移先の中から，教えたい状態へ学習者を導くように次の状態を選ぶ．
本研究	元のゲームの行動価値に基づき，起こりうる遷移先の出現確率を偏らせる．報酬は変えない．

遷移先を操作する教示では，既存の研究でも本研究でも，学習者が経験する状態遷移は元の環境とは異なる．そのため，調整後の環境で学習された行動価値は元のゲームの行動価値から偏り，学習される戦略に影響する可能性がある．そこで本稿では，Q 学習エージェントを調整した環境で学習させたうえで，学習後の戦略を調整のない元の環境で評価するとともに，学習された行動価値の偏りを分析する．

3. ランダム要素調整の提案手法

ランダム要素を調整する手法は主に 3 つの処理から構成される．処理の流れは図 3 に示す．

© 2026 Information Processing Society of Japan

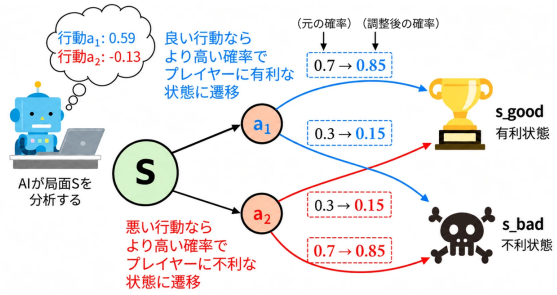


図 3　ランダム要素調整手法のイメージ図

(1) ある局面でプレイヤーが選択した行動を，調整のない通常のゲームにおける行動価値を用いて数値的に評価し，他の行動との行動価値の比較に基づき，良い行動，悪い行動，中立的な行動に分類する．行動価値は，例えば動的計画法による厳密な計算や，強いゲーム AI の評価によって求める．

(2) その後に発生し得るランダム事象を，プレイヤーにとって有利な結果と不利な結果に分類する．有利な結果とは，その結果が起きた後の勝率や期待報酬を高める方向に働くものを指し，不利な結果とはそれらを低める方向に働くものを指す．

(3) 行動の分類とランダム事象の分類に基づき，ゲーム環境内のランダム事象の確率分布を調整する．良い行動の後には有利な結果が，悪い行動の後には不利な結果が起こりやすくなるようにする．中立的な行動の後には，通常の確率分布を用いる．

本手法では，単に有利・不利な事象を増減させるだけでなく，行動の良し悪しが，プレイヤーにとって納得しやすい形で結果に表れることを目指す．本稿で用いる 2 つのゲームでは，行動の直後に生じるランダム事象がその行動の結果を大きく左右するため，結果と自分の選択とのつながりを認識しやすい．ただし，行動の成否が何手も後になって明らかになるゲームも多く，一般的な手法を考える際には，このつながりを定量化し，それに応じて調整する方法を検討する予定である．

4.　ブラックジャックにおける実験（実験 1）

4.1　実験 1 の概要

本研究は，ランダム要素を調整して，プレイヤーの良い行動には良い結果を，悪い行動には悪い結果を与え，プレイヤーの不満低減と上達支援を目指す．本稿では被験者実験に先立ち，簡単なゲームで具体的な手法を示すとともに，強化学習エージェントによる予備実験を行う．本章では，ランダム要素を調整したブラックジャックでエージェントを学習させ，学習後の戦略をランダム要素調整のない元のゲームで評価する．ブラックジャックは構造が比較的単純で，動的計画法により行動価値を厳密に求められるため，行動の良し悪しと学習の結果を正確に評価できる．

4.2　対象ゲームと実験環境の説明

ブラックジャックは，プレイヤーとディーラーが手札の合計を 21 に近づけるカードゲームであり，21 を超えると敗北する．基本的な進行については，脚注に示す解説を参照されたい[*1]．

本実験では，簡易なルールを採用し，プレイヤーが取れる行動を hit，stand，double の基本的な 3 種類とした．

- hit: カードを 1 枚引く行動
- stand: それ以上カードを引かずに勝負をする行動
- double: 最初の 2 枚の状態でのみ選択でき，賭けを 2 倍にしてカードを 1 枚だけ引き，その後は自動的に stand する行動

カードは無限デッキから配られ，A〜K の各カードが常に 1/13 の確率で出る（10 と絵札は 10 点として扱う）．ディーラーは，プレイヤーの手札によらず，自分の合計が 17 以上（A を 11 と数えた 17 を含む）なら必ず stand し，16 以下なら必ず hit する．報酬は勝ち，引き分け，負けに対して通常時はそれぞれ $+1, 0, -1$，double 時をその 2 倍とした．

4.3　実験の流れ

4.3.1　ゲームの完全解析と行動の評価

まず，動的計画法により元のブラックジャック環境における各状態の hit，stand，double の行動価値（Q 値，最適戦略下の期待報酬）を算出した．状態は，プレイヤーの手札の合計，ディーラーの表のカード，A を 11 と数えているかどうか，double の可否の組であり，行動を選ぶ状態は 490 個である．手札の合計では，A を通常 11 と数え，バーストする場合は 1 と数える（例えば A，3 は 14）．以下，この行動価値を $Q^*(s,a)$ と書き，Q^* が最大となる行動を最適行動，各状態で最適行動を選ぶ戦略を Oracle と呼ぶ．プレイヤーが行動を選ぶ度に，その行動を他の行動との Q^* の比較に基づき，良い行動 A_{good}，悪い行動 A_{bad}，中立的な行動 A_{neutral} と分類する．double は賭けが 2 倍になるため，他の行動との価値の差が極端になりやすい．そこで，hit と stand の比較を基準とし，double はその範囲との比較により別に分類した．

- hit，stand: 両者のうち Q^* が高い方を良い行動，低い方を悪い行動とする（等しい場合は中立的な行動）．
- double: Q^* が hit と stand の高い方以上であれば良い行動，低い方以下であれば悪い行動，その間であれば中立的な行動とする．

ただし，hit と stand に明らかな差がある場合と，ほとんど差がない場合を同様に扱い，良い結果や悪い結果に誘導することはやや乱暴な設定であり，今後の検討課題である．

4.3.2　ランダム事象の分類

同じ計算結果を用いて，次にプレイヤーまたはディー

[*1] 株式会社ブライト．“ブラックジャック”．（参照 2026-10-23）casinocafe.jp/casinovenus/game/blackjack

© 2026 Information Processing Society of Japan

ラーに配布されるカードが期待報酬に与える影響を評価し，期待報酬が高くなるカードを有利なカード群 G_{good}，低くなるカードを不利なカード群 G_{bad} として分類した．具体的には，行動後に引く各カード c の効用 $u(c)$ を元のゲームの値で次のように定め，$u(c) > 0$ のカードを有利なカード群，$u(c) < 0$ のカードを不利なカード群とする．

- プレイヤーが hit または double で引くカード:

$$u(c) = B(s,a,c) - V^*(s), \quad V^*(s) = \max_{a'} Q^*(s,a') \tag{1}$$

ここで $B(s,a,c)$ は，状態 s で行動 a を選んでカード c を引いた後の期待報酬である．hit ではその後も最適行動を取るとし，double では引いた 1 枚で勝負する（賭けは 2 倍）．

- プレイヤーの行動終了後に開かれるディーラーの伏せ札と，ディーラーの hit 時の補充札: 同様に，そのカードが出た後のプレイヤーの期待報酬の増減で分類する．

4.3.3 ランダム事象の調整

調整する際には，良い行動なら有利なカード群から，悪い行動なら不利なカード群から，次のカードを元の確率に比例して抽選する．ここでは 10，J，Q，K をそれぞれ別のカードとして扱うため，10 点のカードは組の中でも他のカードの 4 倍の重みを持つ．中立的な行動では，通常の無限デッキ分布を用いてカードを抽選するようにした．また，調整の強さが学習に与える影響を検証するために，本実験ではランダム事象の調整頻度を操作する調整適用率 α を導入した．調整対象の各抽選で α に基づき調整を適用するかを確率的に決定し，非適用時は通常の無限デッキ分布からカードを抽選する．つまり，直前の行動が A_{good} の要素なら $G = G_{\text{good}}$，A_{bad} の要素なら $G = G_{\text{bad}}$ とし，通常の分布を p_0 とすると，カード c が引かれる確率は

$$p_\alpha(c) = (1-\alpha)\, p_0(c) + \alpha \frac{p_0(c)\, \mathbf{1}[c \in G]}{\sum_{c' \in G} p_0(c')} \tag{2}$$

となる．ここで $\mathbf{1}[\cdot]$ は条件が成り立つとき 1，それ以外は 0 をとる関数であり，$\sum_{c' \in G} p_0(c')$ は G に含まれるカードの元の確率の合計である．調整の対象とする抽選と，用いる行動の分類は次の通りである．

- hit または double で引くカード: その行動の分類
- ディーラーの伏せ札と補充するカード: 最後の行動（stand，double，または合計が 21 になった hit）の分類

プレイヤーの最初の 2 枚とディーラーの表のカードは調整しない．調整するのはカード配布の確率分布のみで，報酬の定義は変えない．ただし，カードの出方が変わるため，各行動の期待報酬は元のゲームとは異なる．調整適用率 α は 0%から 100%まで 5%刻みの 21 段階とし，$\alpha = 0$（調整なし）を baseline とする．

4.3.4 学習と評価

学習には表形式の Q 学習エージェントを用い，学習アルゴリズムは調整の有無によらず同じとした．Q 値の初期値は 0，割引率は 1 とし，行動毎に Q 値を更新する．行動は ε-greedy で選び，探索率 ε は最初の 20 万局を 1.0，50 万局までを 0.5，その後を 0.1 とした．学習率は最初の 50 万局を 0.05，その後を 0.01 とした．各条件で 5 つの乱数シードについて 100 万局ずつ学習させた．

評価時は調整を行わず，元のブラックジャック環境で学習済み戦略を評価した．評価指標は Song[6] にならい，学習された戦略と最適戦略（Oracle）の一致率（Oracle matching rate）を用いた．一致率は，490 状態それぞれで学習された戦略の貪欲行動が最適行動に含まれるかを調べ，その割合として求める．一致率は 2,500 局ごとに評価した．

4.4 強化学習の結果とその評価

4.4.1 学習効果

図 4 に学習中の一致率の推移を示す．横軸は学習した局数，縦軸は一致率であり，baseline と $\alpha = 20\%$，50%，100% の 4 条件を示す．調整したいずれの条件でも，baseline より学習の初期から一致率が高く，より早く最適戦略に近づいた．baseline は 100 万局の学習でも最適戦略に届かないのに対し，$\alpha = 50\%$ では学習の比較的早い段階でほぼ最適戦略に達した．一方，$\alpha = 100\%$ は早い段階で頭打ちとなり，最終的な一致率は 20%や 50%より低かった．

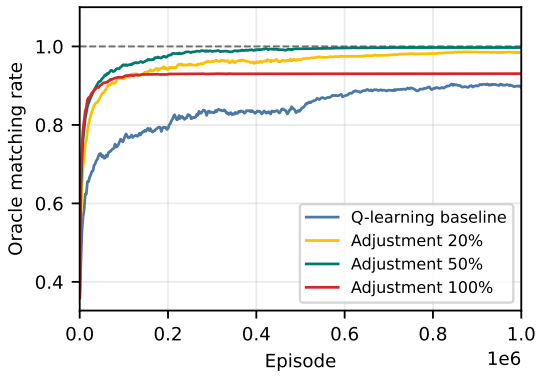


図 4 一致率の推移（5 シードの平均）

次に，学習後の戦略を元のゲームで 1 局あたりの期待報酬により評価した（図 5）．横軸は α，縦軸は学習後の戦略の期待報酬（賭け金を 1 とする）であり，破線は Oracle の値である．期待報酬は動的計画法により求めた．調整した全条件で baseline を上回り，α が 30%から 90%の範囲では Oracle とほぼ同じ値となった．$\alpha = 20\%$ では一致率は高いものの，最初の 2 枚で double すべき状態の一部を hit と学習しており，その分だけ期待報酬が低い．一方，$\alpha = 100\%$ で誤りとなる 34 状態は，元のゲームでは 1 回あたりの損失は大きいものの訪問が稀であり，期待報酬の低下は小さい．

以上から，ブラックジャックでは，弱い調整でも学習は促進され，中程度の調整では元のゲームでも Oracle に近い戦略が得られたが，最も強い調整では一致率が低下した．

4.4.2 調整がゲーム環境に与える影響

本実験の調整では，プレイヤーやディーラーが引くカー

© 2026 Information Processing Society of Japan

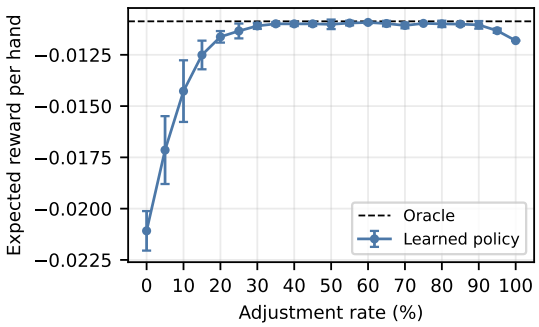


図 5 元のゲームにおける学習後の戦略の期待報酬（5 シードの平均．誤差棒は標準偏差，破線は Oracle）

ドに，元の確率（A〜9 は 1/13，10 点のカードは 4/13）とは異なる確率を割り当てる．式 (2) によれば，α が大きいほど，また G に含まれるカードの元の確率の合計が小さいほど，G 内のカードの確率は大きく引き上げられる．例えば，プレイヤーが良い行動を取ったが，A を引かなければ負けるような状況があれば，$\alpha = 100\%$ では必ず A が引かれる．このような極端な調整は，プレイヤーに調整を疑わせたり，状態の良さを誤って見積もらせたりする原因になりうる．そこで本節では，実際にどの程度の確率の変化が生じているかを調べる．

確率の変化の大きさは，調整対象となる各抽選について，G に含まれるカードの確率の倍率 $p_\alpha(c)/p_0(c)$ で表す．学習後の戦略で調整後の環境を各シード 10 万局対局させ，現れた抽選についてこの倍率の中央値と，倍率が 2 以上となった抽選の割合を求めた．これらを Oracle との一致率，及び 490 状態のうち学習中に 1 回以上訪問された状態の数と併せて，21 段階から代表的な 5 条件を表 3 に示す．調整が強いほど確率の変化は大きくなるが，一致率は中程度の調整でほぼ最大に達し，それ以上調整を強めても向上しない．すなわち，中程度の調整は比較的小さな確率の変化で大きな学習効果を得ている．なお，組に含まれるカードが 1〜2 種類しかないような極端な調整は，いずれの条件でも抽選全体の約 3%にとどまった．

表 3 一致率，出現確率の倍率，学習中の訪問状態数（5 シード）

α	一致率	倍率の中央値	倍率 2 倍以上	訪問された状態
0%	89.5%	N/A	N/A	490
20%	98.5%	1.17	0.01%	490
50%	99.7%	1.36	8.4%	490
90%	99.3%	1.64	32.0%	489.0
100%	93.1%	1.71	33.0%	456

一方，調整を強めると元々稀な状態がさらに現れにくくなる．$\alpha = 90\%$ では，構造上は全ての状態に到達できるが，一部のシードでは 100 万局の間に一度も訪問されない状態が残った．$\alpha = 100\%$ では組に含まれないカードの確率が 0 になるため，hit の後に現れるハード 6，7 などの 34 状態には到達できず，2.3 節で予想した通り，それらの状態は学習されなかった．

4.4.3 行動価値の偏り

図 6 に学習された行動価値と元のゲームの行動価値 Q^* との差の平均を，行動の分類ごとに示す．横軸は α，縦軸は 1 局あたりの期待報酬の差である．良い行動の価値は高く，悪い行動の価値は低く学習され，その差は調整が強いほど大きい．行動間の価値の差が広がることで，少ない経験でも正しい行動の順位を学びやすくなったと考える．ただし，学習された価値は元のゲームの価値とは一致せず，良い行動は過大に，悪い行動は過小に評価される．このような価値の偏りは，プレイヤーに状態や行動の良さを誤って見積もらせる原因になりうるため望ましくない．したがって，調整適用率 α は，学習を促す効果が得られる程度に抑えるべきである．

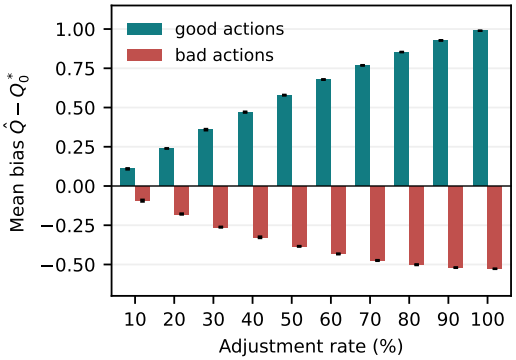


図 6 行動分類別の Q 値の偏り（学習値と元の Q*との差．全条件で訪問された 456 状態，5 シード平均）

5. 3 × 3 盤面の 2048 における実験（実験 2）

5.1 実験 2 の概要

実験 1 では，ブラックジャックにおいてランダム要素調整が良い行動の学習を促すことを確認した．実験 2 では，より大規模なゲームで同様の調整を適用しても，効果が得られるかを検証する．対象は 2048[*2] を 3 × 3 の盤面に縮小したミニ 2048 を用いた．ミニ 2048 は約 4,900 万の状態を持つため，表形式の学習は難しいが，山下ら [7] が完全解析しており，各行動の厳密な行動価値を利用できる．

5.2 対象ゲームと実験環境の説明

2048 は，1 回の操作で盤面上の全タイルを上下左右のいずれかの方向に可能な限り移動させ，同じ数字のタイルを合体させていくゲームである．合体すると 2 枚のタイルが元の 2 倍の数字を持つ 1 枚のタイルになり，1 マス分の空きが生じる．また，合体後のタイルの数字が得点に加わる．各行動の後には，空のマス 1 つに新しいタイルが無作為に現れる（2 が 90%，4 が 10%）．盤面が変化しない方向には動かせず，どの方向にも動かせなくなるとゲームは終了する．ミニ 2048 は，盤面が 3 × 3 である点以外は通常

[*2] Gabriele Cirulli. “2048”.（参照 2026-10-23）play2048.co

© 2026 Information Processing Society of Japan

の 2048 と同じルールである．実験 2 では，各行動で得られる得点を報酬とした．

5.3 実験の流れ

以下では実験 1 との違いを中心に述べる．

5.3.1 行動の評価

完全解析の結果から，状態 s で行動 a を選んだときの行動価値を $Q^*(s,a) = r(s,a) + V^*(s^a)$ として求めた．ここで，$r(s,a)$ は合体による得点，s^a は行動の後で新しいタイルが現れる前の盤面，V^* は最適戦略の下での期待得点である．ミニ 2048 では選べる行動が上下左右の 4 つあり，行動間の価値の差も局面によって大きく異なるため，実験 1 のように hit と stand の比較に基づく分類は使えない．そこで，各行動の損失 $l(a) = \max_b Q^*(s,b) - Q^*(s,a)$ を求め，選べる全ての行動（最善の行動を含む）の損失の中央値 m（行動の数が偶数のときは中央の 2 つの平均）を基準として，$l(a) \leq m/2$ の行動を良い行動，$l(a) > m$ の行動を悪い行動，それ以外を中立的な行動とした．この基準では，最も悪い行動の損失がどれだけ大きくても分類は変わらない．なお，1 回の合体で得られる最小の得点である 4 点未満の損失は 0 と見なし，全ての行動の価値の差が 4 点未満の局面では，どの行動も中立的な行動とした．なお，これらの閾値（$m/2$，m，4 点）は暫定的に定めたものであり，良い行動・悪い行動の基準には，人間プレイヤーの実力帯ごとの平均的な損失を用いることも一案である．

5.3.2 ランダム事象の分類と調整

ランダム事象は，各行動の後に現れる新しいタイル（現れるマスと数字の組）である．その一つ o の効用を，元のゲームの値を用いて以下のように定めた．

$$u(o) = V^*(s^a_o) - V^*(s^a) \qquad (3)$$

ここで s^a_o は，s^a に o が現れた後の盤面である．行動の評価と同じく 4 点を差がないと見なす幅とし，$u(o) > 4$ のものを有利な結果，$u(o) < -4$ のものを不利な結果とした．

調整は実験 1 と同じく，良い行動の後には有利な結果，悪い行動の後には不利な結果の組 G から，調整適用率 α の確率で次のタイルを選ぶ．式 (2) と同じく，組 G 内の候補を元の出現確率 p_0 に比例して抽選し，候補間の相対的な重みを保つ．中立的な行動や該当結果がない場合は元の分布に従い，最初の 2 枚は調整しない．

5.3.3 学習と評価

状態数が多いため，学習には表形式ではなく，2048 で広く用いられている N タプルネットワーク [8] で行動価値を近似する Q 学習エージェントを用いた．N タプルネットワークは，盤面上の決まったマスの組（タプル）について，そのマスの数字の並びごとの値を表として持ち，それらの和で価値を表す．実験 2 では，3 つの行，3 つの列，及び 4 つの 2×2 の正方形の計 10 個のタプルを各行動で用いた．報酬と更新式は通常のまま，各行動の Q 値を更新する．評価時には調整を行わず，学習済み戦略に元のミニ 2048 を 500 局プレイさせて，平均得点を評価指標とした．

5.4 強化学習の結果とその評価

調整適用率 α は 0%（baseline），20%，50%，100%の 4 条件とし，各条件で 5 つの乱数シードについて 50 万局ずつ学習させた．学習率は 0.005 とした．行動は常に行動価値が最大のものを選び，ε-greedy などの探索は行わなかった．Szubert ら [8] も明示的な探索を用いず，2048 は環境自体が確率的で十分に多様な経験が得られるため，ε-greedy による探索を加えても性能は向上しなかったと報告している．

5.4.1 学習効果

図 7 に，元のミニ 2048 における平均得点の推移を示す．横軸は学習した局数，縦軸は評価時の平均得点である．$\alpha = 20\%$ と 50%では，5 万局頃以降は baseline より得点が高く，5 つのシードのいずれでも baseline を上回った（最終的な得点は表 4）．50%は 20%よりやや高いが，その差はシードによって異なり，明確とはいえない．一方，$\alpha = 100\%$ では学習の全期間を通じて baseline を下回った．実験 1 では $\alpha = 100\%$ でも baseline を上回ったのに対し，実験 2 では強すぎる調整がかえって学習を妨げた．

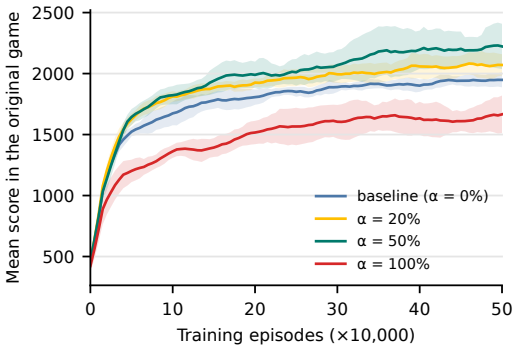


図 7 元のミニ 2048 における平均得点の推移（5 シードの平均．網掛けは± 1 標準偏差）

5.4.2 調整がゲーム環境に与える影響

表 4 に，学習の終盤（45 万〜50 万局）における調整頻度と大きさを示す．実験 1 ではカードの確率の倍率で調整の大きさを表したが，ミニ 2048 では起こりうる結果が多く，その影響の大きさも局面によって大きく異なるため，期待得点の変化 Δ で表す．Δ は，1 回の調整によって次の局面の期待得点（V^* で評価）が通常の出現と比べ，どの程度変わったかを表す．調整した手の割合は α とともに増え，いずれの条件でも有利な調整が不利な調整より多かった．また，ゲームの序盤（盤面の数字の合計が小さい局面）ではほとんど調整されず，調整の大半は盤面がある程度埋まったゲームの中盤以降に起きていた．

© 2026 Information Processing Society of Japan

表 4 学習終盤の調整頻度と大きさ（45 万〜50 万局，5 シード全局を集計）

	0%	20%	50%	100%
1 局の手数	156	164	196	318
調整した手の割合	N/A	3.7%	10.3%	27.1%
（有利，不利）		(2.9, 0.8)	(8.4, 1.9)	(22.9, 4.2)
平均 Δ（有利な調整）	N/A	+128	+132	+158
平均 Δ（不利な調整）	N/A	−422	−305	−254
得点（調整後の環境）	1,937	2,106	2,723	5,233
得点（元のゲーム）	1,947	2,073	2,210	1,640

1 回の調整の大きさは有利と不利で非対称であり，不利な調整の平均 Δ の絶対値は有利な調整より大きかった．ただし，不利な調整の Δ の中央値はいずれの条件でも −32〜−64 点と小さく，平均を大きくしているのは，盤面がほぼ埋まった局面で起こる極端な調整である．例えば，ある局面で損失 171 点の悪い行動を選んだ結果，空のマスが 1 つ残ったとする．そこに 2 が現れれば期待得点は 172 点上がるが，4 が現れると 1,551 点下がり，不利な調整ではこの 4 が必ず現れる．調整適用率 $\alpha = 20\%$ の条件では，ゲーム中盤で終了する傾向があり，不利な調整が適用される局面では期待得点が大幅に低下する傾向が見られた．一方，$\alpha = 100\%$ では終盤まで進む場合が多く，不利な調整が増えるが，期待得点の損失が小さいため，平均 Δ が小さい．

調整は学習中の得点も大きく変え，$\alpha = 100\%$ では，調整後の環境での得点が元のゲームでの得点の約 3 倍に達した．エージェントは有利な結果が頻繁に得られる調整後の環境に適応したが，その戦略が元のゲームでは通用しなかったと考えられる．$\alpha = 20\%$ ではこの差はほぼなかった．

6. おわりに

6.1 実験のまとめ

本稿では，プレイヤーの行動の良し悪しに応じてランダム事象の確率を調整する手法を提案し，ブラックジャックと 3 × 3 盤面の 2048 において，エージェントを調整環境で学習させた後，元のゲーム環境で評価した．中程度の調整は両ゲームで良い行動の学習を促す可能性が示された．ブラックジャックでは，$\alpha = 30\%$〜90%で学習後の戦略が元のゲームでも Oracle とほぼ同じ期待報酬を得た．

一方，強い調整はブラックジャックで到達不能な状態を生じさせ，2048 では元のゲームで通用しない戦略の学習を招いた．原因は異なるが，両結果は調整を中程度にとどめる必要性を示している．また，ブラックジャックでは良い行動の価値が過大評価され，調整が強いほど偏りが大きかった．人間に適用する際は，元のゲームでの勝率を誤認するおそれも考慮し，調整の強さを検討する必要がある．

6.2 一般化に向けた課題

本稿の調整は，行動を良い・悪い・中立に分類し，有利・不利・通常のランダム事象の組から結果を選ぶ単純なものである．そのため，罰や報酬の大きさは行動の良し悪しの程度に対応しない．例えば 5.4.2 節の例のように，罰の大きさは行動の損失ではなく局面の危険度で決まり，損失のそれほど大きくない行動が良くも悪くも大きく罰されることがある．また，中央値を基準としているため，最善に近い行動が複数ある局面では，損失が僅かな行動でも悪い行動と判定されることがあった．このように，それほど悪くない行動の後に極めて不運な結果が，どちらかと言えば良い行動の後に極めて幸運な結果が生じうる．これはプレイヤーの体験としては不自然で，不満の原因になりうる．一方で，自動車教習用シミュレータのように，あえて僅かな不注意が事故を招くように作って練習に用いている場合もあり，その良し悪しは用途による．

一般化に向けては，次のような方法が考えられる．行動の分類については，2048 で用いた中央値に基づく基準の他に，同程度の実力のプレイヤーがその局面で平均的に生じる損失を基準とする方法がある．調整については，誤りの深刻さに応じて調整の強さを変える方法や，極端な幸運や不運が生じにくいように分布を設計する方法が考えられる．

本稿はエージェントによる検証にとどまり，人間プレイヤーの体験や学習への効果は未検証である．今後は，前述した分類と調整の改良を検討するとともに，被験者実験によって，提案手法がプレイヤー体験の観点から望ましいか，また悪い行動の学習が抑えられるかを確かめる予定である．

参考文献

[1] Hunicke, R.: The Case for Dynamic Difficulty Adjustment in Games, *Proc. ACE '05*, pp. 429–433 (2005).

[2] 野村久光，テンシリリックン シラ，池田 心：不満を抱かせにくいゲーム用擬似乱数列の生成と利用，第 9 回エンターテイメントと認知科学シンポジウム論文集，pp. 9–12 (2015).

[3] Kumari, S., Deterding, S. and Freeman, J.: The Role of Uncertainty in Moment-to-Moment Player Motivation: A Grounded Theory, *Proc. CHI PLAY '19*, pp. 351–363 (2019).

[4] Gupta, P., Biswas, S. and Srivastava, V.: Fostering Human Learning in Sequential Decision-Making: Understanding the Role of Evaluative Feedback, *PLOS ONE*, Vol. 19, No. 5, e0303949 (2024).

[5] Zhang, X., Bharti, S. K., Ma, Y., et al.: The Sample Complexity of Teaching by Reinforcement on Q-Learning, *Proc. AAAI Conf. on Artificial Intelligence*, Vol. 35, No. 12, pp. 10939–10947 (2021).

[6] Song, K.: Evaluating Model-Free Policy Optimization in Masked-Action Environments via an Exact Blackjack Oracle, arXiv:2603.18642 (2026).

[7] 山下修平，金子知適，中屋敷太一：3 × 3 盤面の 2048 の完全解析と強化学習の研究，ゲームプログラミングワークショップ 2022 論文集，pp. 1–8 (2022).

[8] Szubert, M. and Jaśkowski, W.: Temporal Difference Learning of N-tuple Networks for the Game 2048, *Proc. IEEE CIG 2014*, pp. 1–8 (2014).

© 2026 Information Processing Society of Japan