人間の行動原則の制約下で自動獲得されたビデオゲーム COM プレイヤの「人間らしい」振る舞いの主観評価

藤井 叙人[1,2,a)]　佐藤 祐一[1]　若間 弘典[1]　風井 浩志[1,b)]　片寄 晴弘[1,c)]

概要：ビデオゲームエージェント（COM）の振る舞いのデザインにおいて，『強い』COM の自律的獲得は「熟達者に勝つ」という目標を達成しつつある．一方で，獲得された COM の振る舞いは，過度に最適化され機械的に感じるという課題が浮上している．この課題を解決するため，著者らは，『人間の行動原則』を課した強化学習や経路探索により，人間らしい COM を自律的に構成するフレームワークについて提案してきた．しかし，それらの COM が本当に人間らしいと解釈されるかどうかの検証が不十分であった．本論文では，自動獲得された COM の振る舞いについて主観評価実験を実施する．

Evaluating Human-like Video-Game Agents Autonomously Acquired with Biological Constraints

Fujii Nobuto[1,2,a)]　Sato Yuichi[1]　Wakama Hironori[1]　Kazai Koji[1,b)]　Katayose Haruhiro[1,c)]

Abstract: While various systems that have aimed at automatically acquiring behavioral patterns have been proposed and some have successfully obtained stronger patterns than human players, those patterns have looked mechanical. We propose the autonomous acquisition of NPCs' human-like behaviors, which emulate the behaviors of human players. In our previous study, the behaviors are acquired using techniques of reinforcement learning and pathfinding, where *biological constraints* are imposed. In this paper, We evaluated human-like behavioral patterns through subjective assessments, and discuss the possibility of implementing the proposed system.

1. はじめに

エンタテインメント系システムにおけるプレイフィール（プレイ時の感覚や印象）の形成は，ユーザ数や売上に直結する重要な事項である．日常娯楽の一翼を担っているビデオゲーム市場に目を向けると，プレイフィールの形成に大きな影響を及ぼす要因として，ゲーム内に登場するコンピュータ担当のエージェント（=COM）の存在を無視することはできない．ゲームにおけるエンタテインメント性の創発と維持のためには，綿密にレベルデザイン（プレイヤのレベルにあわせた難易度の調整）された，“人間プレイヤを楽しませるための『人間らしい』COM” が必要不可欠である．そのため，ゲームにおける COM の振る舞いのデザインには，長らく，ゲームプログラマによる煩多な作り込みが実施されてきた．

COM の振る舞いのデザインにおける作業負荷の軽減と，機械学習の応用領域としての学術的知見の発見を目的として，国内外で COM の自律的獲得に関する研究が執り行われている [1], [2], [3]．この結果，人間の熟達者をも凌駕する “勝つための『強い』COM” の自律的獲得に至っているが，これらの振る舞いは過度に最適化されており，人間にとっては機械的に映る．強い COM を人間プレイヤの代替として扱った場合，エンタテインメント性が欠落するという問題が浮き彫りになっており，“人間プレイヤを楽しませるための『人間らしい』COM” の自律的獲得に興味が集

[1] 関西学院大学大学院 理工学研究科，Graduate School of Science and Technology, Kwansei Gakuin University
[2] 日本学術振興会特別研究員 DC2，Research Fellow of Japan Society for the Promotion of Science
[a)] nobuto@kwansei.ac.jp
[b)] kazai@kwansei.ac.jp
[c)] katayose@kwansei.ac.jp

まりつつある [4], [5].

人間らしい COM を試作検討する研究として，人間プレイヤの振る舞いを記録し機械学習により模倣する手法 [4]，強い COM に対して恣意的にエラーを導入する手法 [5] などが発表されている．これらの研究では，開発者が意図した「強くない」COM のデザインが可能となっており，レベルデザインの一アプローチとしては有効である．しかし，「どのような振る舞いが人間らしいか」ということ自体が，そもそも形式知化されていないため，開発者の経験（ヒューリスティック）による煩多な作り込み，という枠を出ない．

筆者らの研究では，『人間の行動原則』の条件下での強化学習や経路探索により，人間らしい COM の振る舞いを自律的に獲得するフレームワークについて提案してきた [6]．人間の行動原則として，「身体的な制約：“ゆらぎ”“遅れ”“疲れ”」「生き延びるために必要な欲求：“訓練と挑戦のバランス”」を課したフレームワークにより，アクションゲームの “Infinite Mario Bros.” において，「ためらい」や「余裕」といった感情を想起する振る舞いの自動獲得に成功している．しかし，人間プレイヤにおいて，それらの振る舞いが本当に人間らしいと解釈されるのか，人間らしさを解釈するための基準とは何であるかの検証が不十分であった．そこで，本論文では，自動獲得された COM の振る舞いについて主観評価実験を実施し，人間の行動原則を導入することの妥当性を検証する．

以下，第 2 章で，関連研究を紹介し，第 3 章で，人間の行動原則を導入する意義と，その定義を述べる．第 4 章で，“Infinite Mario Bros.” の仕様と，振る舞い獲得の方法について説明する．第 5 章で，獲得された振る舞いの人間らしさを主観評価実験により検証する．

2. 関連研究

2.1 強い COM を追求した研究

振る舞いを自動的に獲得する手法として，教師データを入力とする事例参照型の手法 [7] と，ゲーム木探索や試行錯誤による非事例参照型の手法 [1], [2] がある．教師あり学習は前者に，経路探索や強化学習は後者に分類される．

教師あり学習は，事前に与えられた大量のデータセットを教師データ（入力データに対して出力されるべきデータの例）とし，有用なルールを学習する手法である．教師あり学習によるアプローチの代表的な研究として，保木は，コンピュータ将棋プログラムである Bonanza を提案している [7]．Bonanza は，プロ棋士の棋譜 6 万局のデータを教師とし，将棋の局面における評価関数を自動学習することで，従来手法よりも良い振る舞いを得ることに成功している．この手法は Bonanza メソッドと呼ばれ，多くのコンピュータ将棋プログラムで採用されている画期的な手法である [3], [8]．将棋のように，強い人間プレイヤの膨大な棋譜データが用意できる場合には，教師あり学習による振る舞い獲得は有効である．

経路探索は，ゲーム木におけるスタートからゴールまでの，最小コストとなる経路を探索する手法である．経路探索によるアプローチの代表的な研究として，Robin は，2009 年の Mario AI Competition において，A*アルゴリズムに基づいた COM を構築し優勝している [1]．Mario AI Competition とは，“Infinite Mario Bros.”（ランダムに生成されるステージを制限時間内に攻略する，「スーパーマリオワールド」のようなアクションゲーム.）を対象とした COM の評価コンテストである [9]．Robin の COM は，マリオや敵の動きを事前に解析し，A*アルゴリズムを用いた経路探索によって，ステージをほぼ最適解で攻略することが可能となっている．

強化学習は，自身の振る舞いの試行錯誤を繰り返すことで最適な振る舞いを獲得する手法である．強化学習によるアプローチの代表的な研究として，藤田らは，カードゲームの Hearts を題材とし，Q 学習を用いて，COM の振る舞い獲得に成功している [2]．巨大な状態空間となること，相手の所持するカードを観測できないこと，4 人対戦のゲームであること，の 3 つを Hearts における学習の困難性と考察している．その上で，解決手法として，パーティクルフィルタによるサンプリング，相手の行動予測器，現在の戦局を評価する状態価値関数，ゲームの特徴に基づく次元圧縮を提案し，困難性の解決を図っている．実験の結果，人間の熟達者よりも優れた振る舞いを得ることに成功している．

これらの手法を用いて獲得された COM は，極めて最適であるが故に，人間にとっては機械的と感じる振る舞いを表出してしまう．そのため，エンタテインメント性の向上という視点に立った場合，人間プレイヤの代替として扱うことは憚られる．ゲーム AI 領域では，人間プレイヤが強い COM に勝てなくなる日も近いと考えられており，人間らしい COM の構築が最重要課題となりつつある．

2.2 人間らしい COM を実装した研究

人間らしい COM を実装した関連研究として，Jacob らは，2012 年の The 2K BotPrize において，大会史上初となる，人間よりも人間らしいと評価される COM の構成に成功している [4]．The 2K BotPrize とは，FPS（一人称視点シューティングゲーム）を対象とした，COM の人間らしさを競う評価コンテストである．人間プレイヤの振る舞いをトレースしたデータベースを基に，人間らしいと思われる振る舞いを決定論的に定義し，ニューラルネットにおける制約として適用している．その結果，対戦相手の人間プレイヤから「人間らしい」と評価される COM の振る舞いが獲得できている．

池田らは，コンピュータ囲碁を対象に，既存の強い COM

に意図的に人間らしいミスをさせることで，手加減と思われない程度の「強くなさ」を実現するための初期的検討を実施している [5]．現在の局面における予測勝率と候補手の選択確率を用いた形勢の制御，楽観派や悲観派といったプレイスタイルによる獲得戦略の分析をしており，ゲームのレベルデザインにおける一アプローチを提案している．

上記の手法は，人間らしいと思われる振る舞いを，開発者が恣意的に定義したものである．そのため，振る舞い獲得における作業負荷の軽減や，フレームワークの汎用性の確保は実現されていない．

3. 人間の行動原則

3.1 人間の行動原則を導入する意義

人間らしい COM の振る舞いに関わるプレイスタイルとそのパラメータは，従来，ゲームプログラマや開発者がヒューリスティックに基づいてアドホックに決定していた．COM の人間らしい振る舞いの特徴が形式知化されていないため，ゲームタイトルや機械学習手法に限定的な作り込みを採用するほかなかった．

従来の手法では，人間プレイヤがゲームをするときに必ず生じる制約や欲求を無視しているために，機械的と感じられる振る舞いが表出していると考えられる．コントローラ操作の反応速度が速すぎる，コントローラのボタンの入力が正確すぎる，常に一定の行動のみを正確に繰り返すといった，人間プレイヤでは実現不可能な振る舞いが表出するケースもある．また，レベルデザインを意識しすぎると，ゲームの途中から急に弱くなる，あからさまなコントローラ操作のミスをするといった，プレイスタイルの統一性が崩壊した振る舞いが表出するケースもある．これらの振る舞いは，「相手がいんちきをしているのでは」，「本当に自分の力で勝ったのか」という疑念を生むため，人間プレイヤのゲームへのモチベーションを削ぐ要因となっている．

本研究では，『人間の行動原則』の条件下での機械学習により，人間らしい COM の振る舞いを自律的に獲得するフレームワークの構築を目指す．人間の行動原則としては「身体的な制約」と「生き延びるために必要な欲求」を機械学習の制約条件として課する．これらは，人間プレイヤがゲームをするときに必ず生じる制約や欲求であり，開発者のヒューリスティックや，ゲームタイトルごとの人間らしさの解析に頼る必要がない．

身体的な制約に関する研究例として，Cabrera らは人間の指先による倒立棒の制御実験を [10]，大平らは人間の直立姿勢の制御実験を実施している [11]．人間の行動制御には「ゆらぎ」「遅れ」「疲れ」といった制約が生じるが，人間は訓練によってこれらの制約を意識的もしくは無意識的に考慮し，安全性とパフォーマンスを両立させる行動制御が獲得できると提唱している．

また，生き延びるために必要な欲求について，Maslow は人間の欲求を 5 段階の階層構造で理論化した「自己実現理論」を提唱している [12]．原始的な欲求に近い階層から順に，1) 生理的欲求，2) 安全の欲求，3) 所属と愛の欲求，4) 承認（尊重）の欲求，5) 自己実現の欲求，と人間の欲求を分類している．そして，「人間は自己実現に向かって絶えず成長する生きものである」という仮定の下，「訓練」による知識の定着や，「挑戦」による不満の解消といった行動の動機は，5) 自己実現の欲求に帰結すると考えられている．

本研究において，人間の行動原則を導入した COM を構成することで，敵に対する「ためらい」や「余裕」，コントローラ操作の「たどたどしさ」，最適な行動を模索する際の「熟慮（試行錯誤）」といった，非合理な振る舞いが表出される可能性がある．さらに，安全性とパフォーマンスを両立した振る舞いとは，「わざとらしさ」や「明らかな弱さ」を露呈しない，「統一性のある強くなさ」が再現されている可能性が高い．その結果，あたかも「臆病なプレイスタイル」や「大胆なプレイスタイル」という戦略をもっているかのような，人間らしい振る舞いを表出する COM が，自律的に構成できると考えられる．

3.2 人間の行動原則の定義

前節で述べた，Cabrera ら [10] や大平ら [11] の「身体的な制約」と，Maslow の自己実現理論 [12] で議論されている「生き延びるために必要な欲求」を考慮し，『人間の行動原則』を「身体的な制約：“ゆらぎ” “遅れ” “疲れ”」，「生き延びるために必要な欲求：“訓練と挑戦のバランス”」として，以下のように定義する．

(1) **センサ系，運動系における「ゆらぎ」**
人間プレイヤは，操作対象や敵オブジェクト等の位置（座標）を正確に観測し認識することは難しく，必ず誤差（ゆらぎ）が生じる（見間違い，操作ミスなど）．そこで，COM が観測する操作対象の現在位置やゲームの局面情報に対し，ガウスノイズを付与することで再現する．

(2) **知覚から運動制御に至る「遅れ」**
人間プレイヤは，ゲームの局面を認識してから，実際に動作するまでに遅れが発生する（眼と手の協応動作における遅延など）．そこで，COM が観測する操作対象の現在位置やゲームの局面情報を，数百ミリ秒過去の情報にすることで再現する．

(3) **キー操作の「疲れ」**
人間プレイヤは，ゲームのコントローラのキー操作を，極めて短時間で何度も，または，長時間連続して実施すると疲れが生じる（ボタン連打，単調な操作の繰り返しなど）．そこで，振る舞いを学習する際に，COM にキー操作変更による負の報酬を与えることで再現する．

(4) **「訓練と挑戦のバランス」**

人間プレイヤは，同じ行動を繰り返す事で「訓練」する一方で，同じ行動の結果に飽きたり，その行動で失敗を繰り返したりすると，飽きや失敗を解消するための新奇な行動に「挑戦」する．そこで，失敗を繰り返しているゲーム局面では，新奇な行動に挑戦する傾向を高め，逆に，失敗をほとんどしないゲーム局面では，同じ行動を繰り返して訓練する傾向を高めることで再現する．

4. 振る舞い獲得フレームワーク

4.1 行動原則を課した強化学習

ビデオゲームにおいては，教師となるプレイデータが大量に用意できないため，非事例参照型の手法である強化学習手法を用いることにする．強化学習手法のなかでも，ゲーム内での形勢を報酬という形で直感的に設定できる Q 学習 [13] を用いる．Q 学習では，最適なルールの獲得として学習が進む点で，ゲームプログラマが利用しやすいというメリットもある．

Q 学習では，ゲームのある局面における最適な行動を以下の式で算出する．

$$argmax_{a_t} Q(s_t, a_t) \tag{1}$$

数式 1 において，t はゲーム開始からの時刻，s_t は時刻 t におけるゲーム局面，a_t は時刻 t において COM が選択する行動，$Q(s_t, a_t)$ は局面 s_t と行動 a_t の組に対する，Q 値とよばれる評価値である．つまり，Q 学習では，局面 s_t において Q 値が最も高くなる行動が最適であると出力される．

また，COM が行動した際に，以下の式で Q 値を更新することにより学習が可能となる．

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha((r + \gamma max_p Q(s_{t+1}, p)) \tag{2}$$

数式 2 において，α は学習率と呼ばれる，Q 値の更新において新たな報酬 r をどれだけ重視するかを示す値，γ は割引率と呼ばれる，0 以上 1 以下の定数である．r は局面 s_t において行動 a_t を選択したことによって得られる報酬である．COM の行動選択手法としては $\epsilon - greedy$ 法を用いる．$\epsilon - greedy$ 法は，$1 - \epsilon$ の確率で Q 値が最大となる行動を選択し，ϵ の確率でランダムに行動を選択する．

ビデオゲームにおいては，時刻 t の扱い方として，リアルタイム性があるゲームではフレーム単位，手番が交互に廻るゲームでは手番単位となる．局面 s_t や行動 a_t が無数に設定できる場合は，学習が実時間で収束するようゲーム特徴を考慮した状態圧縮が必要である．また，報酬 r として，操作対象の進んだ距離，経過時間，局面が遷移する際の評価値（形勢）の増減などを与えることで，COM の振る舞いの自動獲得が可能となる [2], [14].

人間の行動原則の導入に関して，「ゆらぎ」と「遅れ」は，数式 2 の $Q(s_t, a_t)$ の計算の際に，数百ミリ秒過去の位置情報や局面情報にガウスノイズを付与したものを s_t とすることで実現する．「疲れ」は，数式 2 の Q 値の更新の際に，報酬 r にキー操作変更による負の報酬を与えることで実現する（報酬 r の詳細については節 4.4 で述べる）．「訓練と挑戦のバランス」は，ランダム行動選択確率 ϵ の設定において，失敗を繰り返しているゲーム局面 s_t では大きな値を設定することで，新奇な行動に挑戦する傾向を高め，逆に，失敗をほとんどしないゲーム局面 s_t では小さな値を設定し，同じ行動を繰り返して訓練する傾向を高めることで実現する．

4.2 行動原則を課した経路探索

有名な最短経路探索手法である A*アルゴリズムにおいて，人間の行動原則の導入を試みる．節 2.1 で述べたとおり，A*アルゴリズムはアクションゲームにおいて，ほぼ最適解を獲得した実績のある手法である．A*アルゴリズムでは，以下の式によりゲーム木の経路のコストを算出する．

$$f^*(n) = g^*(n) + h^*(n) \tag{3}$$

数式 3 において，$f^*(n)$ はスタートノードから，あるノード n を経由して，ゴールノードに辿り着くまでの経路の推定コストを示す．$f^*(n)$ は二つの推定値の和によって算出される．$g^*(n)$ はスタートノードから現在のノード n までの既知のコストである．$h^*(n)$ はヒューリスティック関数と呼ばれ，現在のノード n からゴールノードまでのコストの推定値である．

人間の行動原則の導入に関して，「ゆらぎ」と「遅れ」は，数百ミリ秒過去のキャラクタの位置情報に対してガウスノイズを付与し，その座標をスタートノードとすることで実現する．「疲れ」は，極めて短時間でのキー操作の変更を禁止することで再現する．「訓練と挑戦のバランス」は，学習フェーズを持たない A*アルゴルズムでは実現不可能であるため対象外とする．

4.3 “Infinite Mario Bros.” の仕様

COM の振る舞いを獲得するにあたり，対象とするゲームとしては，1) 同じ局面を何度も再現できる，2) ゲームの明確な目標が設定できる，かつ，3) ビデオゲームを代表する有名なゲームである必要がある．

本研究では，上記条件を満たし，かつ，ゲームの仕様やゲーム環境パラメータが公開されている，“Infinite Mario Bros.” を対象とし，振る舞いの獲得と，その比較検証，主観評価を実施する．“Infinite Mario Bros.” は，世界的に有名なゲームである “スーパーマリオワールド” を模したアクションゲームであり，そのゲーム画面を図 1 に示す．また，“Infinite Mario Bros.” における仕様は以下のとおりである．

- **ステージの自動生成**

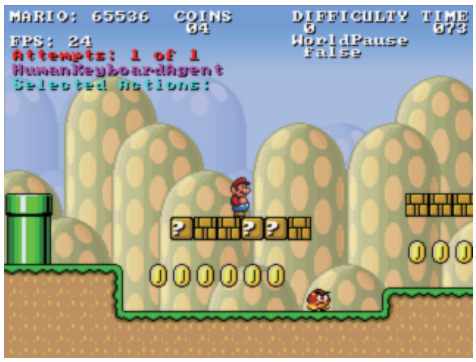

図 1 “Infinite Mario Bros.” のゲーム画面

事前に与えた疑似乱数のシード値に従って無限にステージが生成される．

- **COM の操作キャラクタ（マリオ）**
  COM はマリオ（図 1 中央）を操作する．COM によるマリオの操作はコントローラのキー入力（LEFT, RIGHT, DOWN, SPEED, JUMP）により行う．毎フレームのキーの押下状態により，マリオは対応した行動を行う（毎秒 24 フレームで動作）．
- **敵キャラクタ**
  ステージには数種類の敵が登場し，敵はそれぞれ独自の動作をしている．COM は，これらの敵を避けて進むか，倒して進むかを決定しなければならない．
- **スコアの獲得**
  マリオが死亡する，または，設定された制限時間に達すると攻略は終了し，スコアを獲得する．スコアは Mario AI Competition[9] で規定されている評価関数で計算され，ステージを攻略した距離に応じてスコアが上昇する．
- **COM の観測情報**
  COM は，マリオの座標，マリオの状態，画面内の敵の種類および座標，ステージの地形座標を観測することができる．COM の観測する地形座標は，ステージに配置されているブロックのうち，画面内にある 22 × 22 のブロックの配置座標となる．COM は毎フレーム観測情報を受け取り，マリオの行動制御を行うためのキー入力を返す必要がある．

4.4 “Infinite Mario Bros.” での振る舞い獲得

Q 学習での “Infinite Mario Bros.” の扱い方として，まず，現実的な時間で学習が収束し COM の振る舞いを獲得できるよう，ゲーム局面 s の次元を圧縮する方法を述べる．ゲームの攻略にあたって重要となる情報を削減すると，学習が正常に動作しなくなってしまうことを考慮し，COM の観測できるゲーム局面 s を以下のとおりに圧縮する．

- **マリオを中心に 7 × 7 ブロックの地形と敵配置**
  COM が観測可能な地形座標と敵座標は画面を 22 × 22 ブロックに分割したものである．しかし，1 フレームあたりのマリオの移動距離は小さく，画面内全ての地形座標や敵の配置がマリオの行動に影響することはない．そこで，学習に使用する地形情報と敵の配置は，マリオを中心とした 7 × 7 ブロックとする．これにより，ゲーム局面 s の次元数は大幅に削減される．
- **マリオの進行方向**
  敵や地形との関係性を把握するための重要な要素であるため，COM は 8 方向＋停止の 9 次元としてマリオの進行方向を把握しておく必要がある．
- **「でかマリオ」か「ちびマリオ」か**
  「でかマリオ」でダメージを受けた場合は「ちびマリオ」に変化するだけで攻略を続行できるが，「ちびマリオ」でダメージを受けた場合は死亡となる．より長く攻略を進めるうえで重要な要素であるため，COM はマリオの状態を把握しておく必要がある．
- **マリオが地上にいるか**
  マリオは，地上にいる場合はダッシュやジャンプができるが，空中にいる場合はできない仕様である．マリオが地上にいるかどうかは，行動選択にあたって重要な要素であるため，COM はマリオが地上にいるかどうかを把握しておく必要がある．

表 1 行動の種類とキー入力の組み合わせ

行動の種類	(LEFT,RIGHT,DOWN,JUMP,SPEED)
右に歩く	(OFF,ON,OFF,OFF,OFF)
右に走る	(OFF,ON,OFF,OFF,ON)
右に歩きジャンプ	(OFF,ON,OFF,ON,OFF)
右に走りジャンプ	(OFF,ON,OFF,ON,ON)
左に歩く	(ON,OFF,OFF,OFF,OFF)
左に走る	(ON,OFF,OFF,OFF,ON)
左に歩きジャンプ	(ON,OFF,OFF,ON,OFF)
左に走りジャンプ	(ON,OFF,OFF,ON,ON)
真上にジャンプ	(OFF,OFF,OFF,ON,OFF)
しゃがむ	(OFF,OFF,ON,OFF,OFF)
静止	(OFF,OFF,OFF,OFF,OFF)

次に，Q 学習における選択可能な行動 a の設定方法について述べる．マリオの行動は，コントローラのキー入力によって決定される．マリオの行動制御に影響があるキー入力の組み合わせは 11 パターン存在する．そこで，選択可能な行動 a として表 1 のとおり設定する．

続いて，Q 学習における報酬 r の設定方法について述べる．敵を可能な限り避け，ステージをより早く，より遠くまで攻略するためには，ステージを早く攻略することに対して正の報酬を与え，逆にダメージを受ける，死亡するといった，攻略を阻害する要因に対して負の報酬を与えることが望ましい．また，キー操作変更による疲れを実現するため，キー操作を変更した場合は負の報酬を与える必要がある．そこで，報酬 r を以下のとおり設定する．

$$r = distance + damaged + death + keyPress \qquad (4)$$

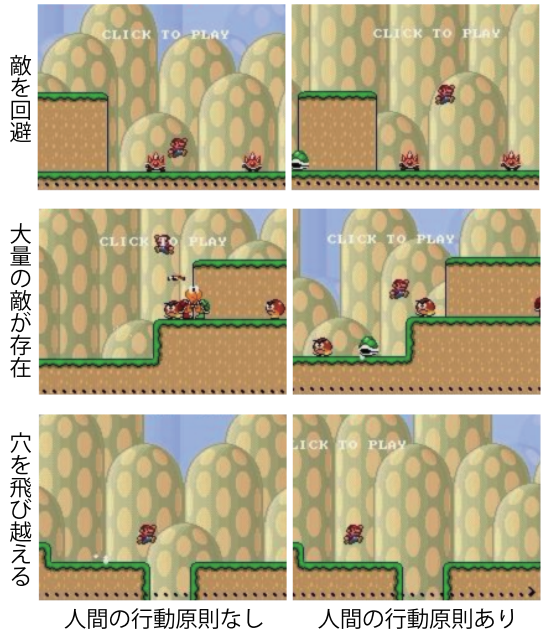

図 2 導入なし（左列）と導入あり（右列）での振る舞いの比較

数式 4 において，*distance* は行動によって進んだ距離であり，そのまま正の報酬とする．*damaged* は行動によってダメージを受けた場合に与える負の報酬，*death* は行動によって死亡した場合に与える負の報酬である．また，*keyPress* は前フレームから行動を変更した場合に与える負の報酬である．予備実験の結果，本研究における *distance* は進んだ距離 ×2.0，*damaged* は-50.0，*death* は-100.0，*keyPress* は-5.0 とした．

最後に，A*アルゴリズムにおけるゲーム木の作成方法と，経路のコスト算出について述べる．節 2.1 で述べた A*アルゴリズムに基づく COM[1] を参考にする．スタートノードを現在のマリオの位置座標（ただし「ゆらぎ」や「遅れ」が付与された座標），ゴールノードを画面の右端とし，マリオが取り得る行動によってゲーム木を作成する．$g^*(n)$ としては，スタートノードから現在ノードまでの時間を，$h^*(n)$ としては，現在ノードから画面の右端に到達するまでの推定時間を算出している（詳細は [1] を参照）．

5. 主観評価実験

5.1 獲得された振る舞い

本研究の振る舞い獲得フレームワークを用いて獲得された COM の振る舞いを図 2 に示す．人間の行動原則の導入の有無によって，表出した振る舞いの特徴に以下のような差異があった．

触れることができない敵を回避する場面（図 2 上段）

- 導入なし（左）: 最小限のジャンプ，かつ，ノンストップで攻略
- 導入あり（右）: 大きくジャンプし，途中で一瞬止まるような行動をしつつ攻略

5 体の敵が段差の上に存在する場面（図 2 中段）

- 導入なし（左）: 正確な行動制御で敵が大量に存在する区間を攻略
- 導入あり（右）: 区間の手前で待機し，安全に進める状態に変化してから攻略

穴を飛び越える場面（図 2 下段）

- 導入なし（左）: 穴に落ちる寸前のところから最小限のジャンプで攻略
- 導入あり（右）: 穴の少し手前から大きくジャンプし余裕を持って攻略

これらの振る舞いの特徴は，Q 学習によって生成された COM（以降，Q 学習エージェント），A*アルゴリズムによって生成された COM（以降，A*エージェント）の双方において共通であった．以上の結果から，人間の行動原則の導入なしでは，パフォーマンスのみを重視しているが，導入有りでは，安全性も考慮した振る舞いが獲得できているといえる．

5.2 実験計画

人間の行動原則を導入したエージェントによって獲得された振る舞いが，本当に人間らしいかどうかを検証するため，20〜24 歳の男女 20 名（男性 13 名，女性 7 名）を対象に主観評価実験を実施した．被験者 20 名における，横スクロール型マリオのプレイ時間の累計は平均 $\mu = 34$ 時間，標準偏差 $\sigma = 29$ 時間であった．そこで，本実験においては，横スクロール型マリオの熟練度を 3 つのグループに分類した．横スクロール型マリオのプレイ時間が 5 時間（$\mu - \sigma$）未満の被験者 4 名を「初級者」（うち，3 名はプレイ時間が 0 時間の初心者），63 時間（$\mu + \sigma$）以上の被験者 2 名を「上級者」，5 時間以上 63 時間未満の被験者 14 名を「中級者」と定義した．

実験手続きは以下の通りである．まず，被験者に「ブロック，アイテム，コイン等は無視して，ステージの先に進め」と教示し，“Infinite Mario Bros.” を 10 回プレイ（1 プレイ 25 秒）させた．次に，プレイ動画を 2 つずつ比較させ「どちらのマリオが人間らしいプレイか」を 7 段階で評価させた．最後に，プレイ動画を 1 つずつ見せ「どのような振る舞いが人間らしい（人間らしくない）と感じたか」を自由記述で回答させた．

実験に使用したプレイ動画を表 2 に示す．本実験では，Q 学習エージェントによるプレイ動画を 3 つ，A*エージェントによるプレイ動画を 2 つ，人間が操作したプレイ動画を上記熟練度を考慮して 3 つ用意した．Q 学習エージェントに関しては，行動原則の導入ありと導入無しの 2 つに加えて，訓練をせず失敗に対する挑戦のみを実施するエージェントも用意した．この Q 学習エージェントにおけるランダム選択確率 ϵ は 0.0，失敗を繰り返しているゲーム局面での ϵ は 0.2 と設定した．人間の操作者に関しては，初級者動画は横スクロール型マリオのプレイ時間が 5 時間の

表 2　プレイ動画のラベルと内容

ラベル	操作者	行動原則	再生時間	スコア
[強化, 無し]	強化学習 (COM)	導入なし	10.62 秒	5448
[強化, 導入]	強化学習 (COM)	導入あり	14.25 秒	4069
[強化, 導入, 挑戦のみ]	強化学習 (COM)	導入あり (挑戦のみ)	15.57 秒	3458
[探索, 無し]	経路探索 (COM)	導入なし	7.29 秒	7926
[探索, 導入]	経路探索 (COM)	導入あり	9.34 秒	3118
[中級者]	中級者 (人間)	−	10.08 秒	6031
[初級者]	初級者 (人間)	−	14.25 秒	3644
[上級者]	上級者 (人間)	−	7.68 秒	7371

人間プレイヤ，中級者動画は 50 時間の人間プレイヤ，上級者動画は 200 時間の人間プレイヤが操作したものである．また，敵，土管，穴といった障害物の有無や，マリオが敵に接触しダメージをうけるシーンが，人間らしさの評価に大きく影響を与えると考えられる．そこで，全ての動画でプレイ区間を統一し，マリオが敵に接触しダメージをうけたプレイ区間は不採用とした．これ以降，プレイ動画をラベル名で表記する．

5.3　分析手法と結果

本実験では，ランダムに表示される 2 つのプレイ動画を比較し，人間らしさについて 7 段階で評価している．統計的分析手法としてシェッフェの一対比較法（中屋の変法）を使用し，分散分析で主効果に対する有意差の有無を確認する．その後，ヤードスティック法によりプレイ動画の嗜好度を一本の直線上にプロットし，動画同士の相対的な関係性と，信頼区間について検討する．本実験では，COM における行動原則の導入の有無による比較，COM と人間プレイヤとの比較に焦点を当てるため，Q 学習エージェントと A*エージェントを分けて分析することとした．

図 3 は，人間らしさに関する相対的嗜好度をプロットしたものである．上の直線は Q 学習エージェントと人間プレイヤの比較，下の直線は A*エージェントと人間プレイヤの比較である．まず，Q 学習エージェント同士の比較結果を述べる．行動原則を導入した [強化, 導入]（相対的嗜好度：0.66）は，行動原則を導入していない [強化, 無し]（相対的嗜好度：0.29）と比較して，人間らしいという結果が得られた．しかしながら，相対的嗜好度の差（0.66 − 0.29 = 0.37）が 95%信頼区間である 0.48 より小さいため，5%水準の有意差は認められなかった．この結果を，以降（差:0.37 < 95%信頼区間:0.48）と表記する．次に，A*エージェント同士の比較結果を述べる．行動原則を導入した [探索, 導入] は，行動原則を導入していない [探索, 無し] と比較して，1%水準で有意に人間らしいという結果が得られた（差:1.35 > 99%信頼区間:0.72）．最後に，COM と人間プレイヤの比較結果を述べる．行動原則を導入した Q 学習エージェント [強化, 導入] は，人間プレイヤの [初級者][中級者][上級者] より人間らしいという結果が得られた．また，行動原則を導入した A*エージェント [探索, 導入] は，人間プレイヤの [初級者][上級者] より人間らしいという結果も得られた．ただし，有意差が認められたのは，[強化, 導入] と [初級者]（差:1.12 > 99%信頼区間:0.58），[強化, 導入] と [上級者]（差:1.33 > 99%信頼区間:0.58），[探索, 導入] と [上級者]（差:0.71 > 95%信頼区間:0.59）のみであった．

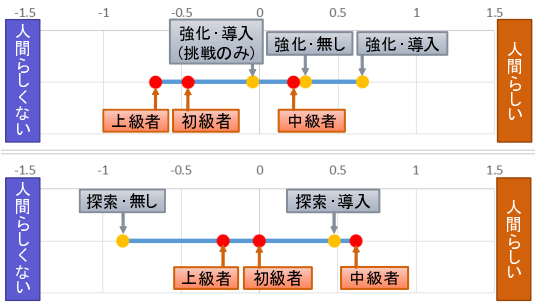

図 3　人間らしさに関する相対的嗜好度

6.　考察

主観評価実験の結果から，人間の行動原則を導入することで，『人間らしい』と解される COM を自律的に構成できることが示された．では，「どのような振る舞いが人間らしいのか」について，主観評価実験の結果（図 3 と自由記述質問の回答）から考察していく．

[強化, 導入] は全動画中で最も人間らしいと評価されている．また，[探索, 導入] は [探索, 無し] と比較すると人間らしい（1%の有意水準で有意差あり）という評価である．自由記述質問では，人間らしいと感じる理由として「敵や穴を飛び越える時に一瞬後ろを向く」，「敵や穴を大きく飛び越える」，「ときどき不必要な行動をとる」という回答があった．この結果から，「ためらい」や「余裕」，「熟慮（試行錯誤）」を感じさせる要素として，『人間の行動原則』を導入することの妥当性が示された．

[上級者] は人間プレイヤの操作であるにもかかわらず，人間らしくないという評価を得ている．また，ほぼ最適解である [探索, 無し] は [上級者] よりもさらに人間らしくない（5%の有意水準で有意差あり）という評価である．自由記述質問では，人間らしくないと感じる理由として「敵や穴をギリギリまで避けない」，「無駄な行動が一切ない」，「動きが一定である」という回答があった．この結果は，「過度に最適化された振る舞いは人間らしくない」ことを意味する．節 2.1 で述べた，強い COM を人間プレイヤの代替として扱うことができない根拠が示された．また，このことから，[強化, 導入] と [強化, 無し] で有意差が認められていない理由も説明できる．[強化, 無し] は [上級者] や [探索, 無し] のスコアに遠く及んでおらず（表 2），最適化

された振る舞いの獲得に至っていないためと考えられる．Q 学習エージェントの改良には，節 4.4 で述べたゲーム局面 s の観測情報を拡張する必要がある．

[強化, 導入, 挑戦のみ] は，行動原則を導入しているにもかかわらず，人間らしくないという評価を得ている．自由記述質問では「段差や土管にぶつかってからジャンプする振る舞いが人間らしくない」という回答があった．この動画は，スコアがかなり低く，その「たどたどしい」振る舞いは，コントローラ操作やゲームルールに慣れていない，あたかもゲーム初心者の操作のようであった．この結果は，「初心者相当の下手すぎる振る舞いは人間らしくない」ことを意味する．また，「訓練と挑戦のバランス」を変化させることで，人間の熟達過程を再現できる可能性が示された．

上記の考察から，ビデオゲームにおいて，人間が人間らしさを解釈するための評価基準を策定できる可能性が示唆された．人間は誰しも，身体的な制約を生得的に持っており，また，生きていくためには適応進化の欲求が必要不可欠である．そのため，これらの制約や欲求が考慮された振る舞いからは，「ためらい」や「余裕」，「熟慮（試行錯誤）」といった感情を想起し，その結果，人間らしい振る舞いであると解釈していると言える．逆に，それらの制約や欲求を無視した「過度に最適化された振る舞い」からは，情緒的な感情を想起することがなく，人間らしさも感じないのであろう．もちろん，「人間らしさの評価基準は被験者間で異なるのではないか」という疑問もある．被験者の横スクロール型マリオの熟練度や，ゲームに対するプレイスタイルにより，被験者を群分けすることで，被験者間の評価基準の差異を検証する必要がある．

7. おわりに

ビデオゲームにおけるユーザエクスペリエンスの向上には，人間らしい COM の実装が必要不可欠であり，その自律的獲得には，従来，ゲームジャンルやゲームタイトルに合った人間らしさの解析が必要であった．本研究では，人間の行動原則を導入することで，人間らしい COM の振る舞いを自動獲得できることが示された．行動原則を課した強化学習や経路探索により，「人間プレイヤがゲームをしている」かのような振る舞いが表出され，また，主観評価実験により，それらの振る舞いが人間プレイヤよりも人間らしいことを示した．人間の行動原則は，開発者のヒューリスティックや，人間らしさの解析に依拠しない要素である．そのため，あるゲーム状況を入力とし，そのゲーム状況で最適な行動を出力する必要があるゲームであれば，ゲームジャンルや振る舞い獲得の手法を問わず，人間らしい COM の振る舞いを獲得できると考えられる

本研究のフレームワークを使用することで，人間らしい COM を実装したいゲームプログラマにとって，1) ヒューリスティックの導入に係る煩多な作業負荷（開発コスト）を削減できる，2) 人間が持つ行動原則であるため生理学的・心理学的知見に基づいて設定できる，3) 様々なゲームジャンル，様々な機械学習手法に対しても，汎用的に導入できる，という 3 つのメリットがある．また，人間らしい COM が実現することで，ゲームをプレイする人間プレイヤにとって，満足感の確保やエンタテインメント性の持続といった，ユーザエクスペリエンスの向上につながると考えられる．今後の展望としては，被験者を群分けし人間らしさの評価基準を特定する，アクションゲーム以外のジャンルにも本研究のフレームワークを適用することを目指す．

参考文献

[1] Togelius, J., Karakovskiy, S. and Baumgarten, R.: The 2009 Mario AI Competition, *Evolutionary Computation (CEC) 2010 IEEE*, pp. 1–8 (2010).

[2] Fujita, H. and Ishii, S.: Model-based reinforcement learning for partially observable games with sampling-based state estimation, *Neural Computation*, Vol. 19, pp. 3051–3087 (2007).

[3] Hoki, K. and Kaneko, T.: The Global Landscape of Objective Functions for the Optimization of Shogi Piece Values with a Game-Tree Search, *Advances in Computer Games 2012, Lecture Notes in Computer Science*, Vol. 7168, pp. 184–195 (2012).

[4] Schrum, J., Karpov, I. V. and Miikkulainen, R.: Human-like Behavior via Neuroevolution of Combat Behavior and Replay of Human Traces, *2011 IEEE Conference CIG' 11*, pp. 329–336 (2011).

[5] 池田心，Viennot, S.：モンテカルロ碁における多様な戦略の演出と形勢の制御〜接待碁 AI に向けて〜, *GPW2012*, pp. 47–54 (2012).

[6] 藤井叙人，佐藤祐一，若間弘典，片寄晴弘：生物の基本原則の導入によるビデオゲーム COM プレイヤの『人間らしい』振る舞いの自動獲得，Vol. 2013-EC-27, No. 16, pp. 1–6 (2013).

[7] 保木邦仁：局面評価の学習を目指した探索結果の最適制御，*GPW2006*, pp. 78–83 (2006).

[8] Sugiyama, T., Obata, T., Hoki, K. and Ito, T.: Optimistic Selection Rule Better Than Majority Voting System, *Computers and Games, Lecture Notes in Computer Science*, Vol. 6515, pp. 166–175 (2011).

[9] J.Togelius, S.Karakovskiy, J.Koutnik and J.Schmidhuber: Super Mario Evolution, *2009 IEEE Conference CIG'09*, pp. 156–161 (2009).

[10] J.L.Cabrera and J.G.Milton: On-Off Intermittency in a Human Balancing Task, *Physical Review Letters*, Vol. 89, No. 15 (2002).

[11] 大平徹，保坂忠明：不安定な状況でのノイズと遅れの役割と制御への考察，交通流のシミュレーションシンポジウム，pp. 19–22 (2004).

[12] Maslow, A. H.: A Theory of Human Motivation, *Psychological Review*, Vol. 50, pp. 370–396 (1943).

[13] Watkins, C.: Learning from Delayed Rewards, *PhD thesis, Cambridge University, Cambridge, England.* (1989).

[14] Patel, P. G., Carver, N. and Rahimi, S.: Tuning Computer Gaming Agents using Q-Learning, pp. 581–588 (2011).