Vol.2013-GI-29 No.7
2013/3/4

プレイヤの効用を学習し行動選択する チームメイト AI の構成

吉谷　慧[†1]　　池田　心[†1]

近年のコンピュータゲームにはコンピュータ AI とチームを組んで一緒に遊べるものも多いが，AI プレイヤは人間プレイヤが望まない行動を取ることがあり，これが人間プレイヤにとってストレスとなる場合がある．これは，この種のゲームが多目的的であり，AI プレイヤが“人間プレイヤはどの目的をどの程度重視しているか”といった人間の価値観（これを効用という）を理解できていないことが原因のひとつとして挙げられる．本研究では人間プレイヤが選択した行動などから人間プレイヤの効用を推定し AI プレイヤの行動選択に活用することで，人間プレイヤの満足度が高い AI プレイヤを生成する．

A Framework of Teammate AI for Learning Human-player Utility and Selecting Own Move

KEI YOSHIGAI[†1]　KOKOLO IKEDA[†1]

In recent years, there are many commercial video games that human-player can team up with some game AIs and play the game together. Many game AIs will often select a move which is worthless for human-player and it will give some kinds of stress for human-player. One reason for this is game AIs don’t understand human-player’s utility. In this paper, we propose a method to estimate human-player’s utility from his/her selection, and build a game AI with the method.

1. はじめに

ゲーム AI 研究の目標のひとつとして「人間を楽しませるゲーム AI」を作ることが挙げられる．近年までのゲーム AI の研究は“強さ”に焦点をあてたものが多く，既にオセロやチェス，将棋といったボードゲームにおいて，ゲーム AI の強さは人間のプロレベルに達している．チームとしての“強さ”に焦点をあてた研究としては，Sander,B.らの研究[1]で，QuakeIII のゲーム AI を動的に敵チームに適応させることで強いチームを生成する手法が提案されている．これら，ゲーム AI の“強さ”に焦点をあてた研究は，人間プレイヤの「敵」として人間を楽しませるための研究といえる．

Aswin,T.A.らの研究[2]では変化していくゲームの難易度に合わせて,チームメイトである Non Player Character のサポート行動をプレイヤに適応させる試みがなされている．また，Kenneth O.S.らの研究[3]では，ゲームプレイ中のようなリアルタイムでも，プレイヤとのやりとりを通じてゲームキャラクタが弱点を克服することで，ゲームの面白さを保つ試みがなされており，人間がゲーム AI を訓練するという新たなゲームジャンルを提唱している．これらは人間の「仲間」として，もしくは“強さ”だけでなく別の要素にも着目し，人間を楽しませるための研究といえる．

しかし,「仲間」として人間を楽しませる AI の研究は「敵」としての研究と比べるとまだ少なく，また“強さ”以外の要素に焦点をあてた研究も少ない．特に日本での人気の高いコマンド形式 RPG（ドラゴンクエスト等）ではこのような技術の需要は大きい．

2. 目的

本研究の目的は「人間が一緒にプレイして“楽しい（ストレスや不満が少ない）”，“仲間”」の AI プレイヤを生成することである．近年市販されているコンピュータゲームではコンピュータ AI が操作するキャラクタとチームを組んで遊べるものも多いが，しばしば AI プレイヤは人間プレイヤが意図しない（望んでいない）行動を取ることがあり，これが人間プレイヤにとってストレスとなる場合がある．

人間プレイヤが AI プレイヤに望む行動と，AI プレイヤが実際に選択する行動が異なってしまう原因を本論文では以下 4 つに分野する．

① 人間プレイヤと AI プレイヤの効用（何を目指して行動しているか，またはその重み）が違うため，それぞれにとって最適な行動が異なるから
② それぞれの効用が同じだとしても，人間プレイヤの情報処理能力が不十分であり最適な行動が分かっておらず，“間違った”不満を持っているため
③ それぞれの効用が同じだとしても，AI プレイヤの探索などが不十分で最適な行動が分かっていないため
④ 本質的に混合戦略が必要なゲームであり,「一定確率で期待と違ってもしょうがない」ため

本研究では①の理由に注目し，AI プレイヤに人間プレイヤ

†1 北陸先端科学技術大学院大学
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY

ⓒ 2013 Information Processing Society of Japan

の効用を学習させることで，行動選択の差異による人間プレイヤへのストレスを軽減したい．

3. 関連研究

協力ゲーム・非協力ゲームにおける戦略はゲーム理論の枠組で論じられることが多い. コマンド形式 RPG の多くは同時着手性をもち，仲間との間には利得を共有する協力ゲームの関係がある．

ただし，こういったゲームの利得は確定的なものではなくランダム性に影響され，リスクをどの程度重く見るかなどを扱うためには効用の考え方を導入する必要がある．

- 効用[4]

  効用とは個々人が持つ，物事への価値観のことである．人間の価値観を定量化することを目的とした効用理論の歴史は，D.ベルヌーイにさかのぼる．人々の行動原理として，ベルヌーイは「人々は期待値を最大化して行動しているのではなく，期待効用を最大化している」と説明し，お金に対する効用関数として対数関数を提案した．これが効用関数の発祥である．

- プロスペクト理論[5]

  Kahneman は，期待効用に基づく選択の矛盾を数多くの実験結果によって指摘したうえで，期待効用値理論に代わる理論として，プロスペクト理論を提案した．人は，ほぼ確実な収益に対しては，実際の確率よりも低い確率を考えがちであり，これを確実性効果と呼んでいる．人は収益の領域ではリスク回避的な確実性効果が強く表れるが，損失の領域では不確実なものを好むようになる．

4. 人間の思考について

4.1 人間の状態評価に関する仮説

人間は“状態を構成する要素”それぞれに対して好ましさを持っていて，各要素を総合的に評価した結果がその状態の良さとなる，と仮定する．

“状態を構成する要素”とは，人間プレイヤや仲間の AI プレイヤ，もしくは敵対プレイヤなどの行動選択によって影響されうるゲーム中のパラメータのことである．ここで“状態を構成する要素”について，一般的なターン制の RPG ゲームにおける戦闘シーンを例に説明する．この種のゲームでは戦闘に参加している敵・味方キャラクタが持つパラメータとして HP（キャラクタの体力のこと．この数値が 0 になると行動できなくなる）や MP（キャラクタが一部の特殊な行動を取るために必要となる数値）などが設定されている場合が多く, これらが“状態を構成する要素”となりうる．また他にも，戦闘に参加しているキャラクタそれぞれの生死や，戦闘の経過時間（経過ターン数）なども，“状態を構成する要素”となりうる．

ひとつ気を付けなければならないのは，人間の持つ効用は個人個人で違うという点である．同じ状態を評価したとしても個人個人の評価は違ってくる場合がある．例えば図 1 のような状態を複数の人が評価した場合，ある人は 2 人共 HP が半分以下という理由で状態を「悪い」と判断するかもしれないが，別の人は MP が残っているので「普通」と判断する可能性がある．

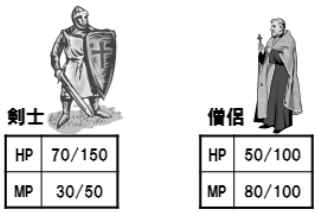

図 1　一般的な RPG ゲームにおける状態の一例

4.2 人間の行動決定に関する仮説

人間は複数の行動選択肢がある場合，それぞれの選択肢について行動した後の“未来の状態”を予想・評価して行動選択をしている．

人間が予想する“未来の状態”がどの程度先の状態かは，個人差やゲームの種類によって様々である．ターン制のゲームならば，1 ターン先の状態や数ターン先の状態，もしくは戦闘終了時の状態などが，人間が予想する“未来の状態”の候補として挙げられる．α β法は前者の，モンテカルロ方は後者の代表的な探索方法である．

図 2 は二人同時着手確定ゲームの状態遷移図と，遷移後の各状態に対する人間の評価を表にしたものである．人間プレイヤの行動選択肢がa_1とa_2，AI プレイヤの行動選択肢がb_1とb_2である．なお，評価表はあくまで人間プレイヤにとっての評価であり, AI プレイヤにとっての評価表が同じとは限らない．状態は両プレイヤの選択の組合せによって確率的に遷移する．

人間プレイヤが図 2 のように評価した場合，実際に選択する行動は AI プレイヤの立場や個人差で違ってくるであろうが（AI プレイヤが仲間の場合，AI プレイヤが賢いと思っている人はa_1を選ぶかもしれないし, AI プレイヤを信頼していない人はリスク回避的にa_2を選ぶかもしれない）いずれにせよ“未来の状態”を予想・評価して行動選択をしている．

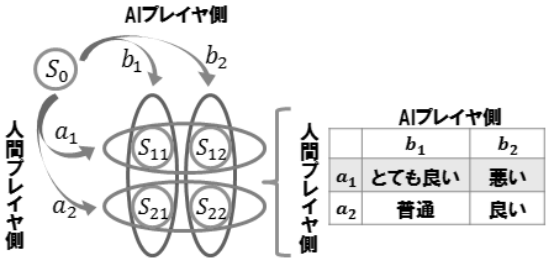

図 2　同時着手確定ゲームの状態遷移図と人間の評価例

4.3 人間のストレスとなりうる AI プレイヤの行動選択

人間が大きなストレスを感じるパターンのひとつとして，「合理的と思えない」「選択理由が分からない」行動を仲間の AI プレイヤが選択したときが挙げられる．AI プレイヤの選択によって評価の低い状態に遷移した場合も，確かに人間プレイヤはストレスを感じるであろうが，それ以上に「選択理由が分からない」行動を AI プレイヤが選択した場合の方が大きなストレスを与える可能性がある．

表 1 は，ある状態で人間プレイヤが“未来の状態”を評価し，その良さを数値化して表にしたものである．人間プレイヤの行動選択肢がa_1～a_3，AI プレイヤの行動選択肢がb_1～b_3である．

人間プレイヤの評価では(a_1, b_3)の組合せの評価が一番高いので，AI プレイヤがb_3の行動を選択する理由を人間プレイヤは理解できる．また，(a_2, b_2)の組合せも(a_1, b_3)程ではないが評価は高いので，AI プレイヤがb_2を選んだ場合も人間プレイヤはその選択に理解を示せるはずである（2 章④を参照）．人間プレイヤの選択次第では(a_2, b_3)や(a_1, b_2)といった評価の低い状態に遷移する可能性もあるが，AI プレイヤの選択に理由付けができるので，ある程度ストレスが緩和されることが期待できる．しかし AI プレイヤがb_1の行動を選択した場合，人間プレイヤにとって大きなストレスとなる可能性がある．これは，b_1の選択肢はb_2やb_3の選択肢に対して優位性が少なく選択する理由として弱いことや，唯一評価が勝っている(a_3, b_1)の組合せも，人間プレイヤにとってa_3はa_2の完全な被支配戦略であり，そもそもa_3の選択肢を考慮に入れること自体があり得ないと思えるからである．

このように納得できる評価の低い状態に遷移したうえ，AI プレイヤの行動選択に理由が見つからない場合，人間プレイヤは大きなストレスを感じる可能性が高い（2 章①を参照）．

人間プレイヤが不快に思わない行動を AI プレイヤが選択するためには，人間の効用を理解し，明らかにその効用に反するような選択肢を選ばないことが大切である．

	b_1	b_2	b_3
a_1	0.2	0.5	**0.8**
a_2	0.4	0.7	0.4
a_3	0.3	0.2	0.2

表 1　人間による“未来の状態”の評価表の一例

5. 本研究で扱うゲームの概要

本研究では独自にコマンド形式のターン制ゲームを設定し，提案手法の実装や実験を行う．このゲームは Markov Decision Process で記述できる多人数同時着手不確定ゲームである．ゲームの各種設定は以降で説明する．

5.1 キャラクタの持つパラメータ

ゲームに参加するキャラクタには以下の通り 4 つのパラメータを設定する．キャラクタを操作するプレイヤは，自身が操作しているキャラクタのパラメータは勿論，敵・味方全てのキャラクタのパラメータを知ることができる．

- 体力（HP）
  体力が 0 になると，そのキャラクタは行動不能となる．チーム全員の体力が 0 になった場合，そのチームは戦闘に敗北する．可変値．
- 精神力（MP）
  一部の術技を選択するために必要な数値で，対象となる術技を使用すると値が消費される．可変値．
- 攻撃力
  攻撃をした際，相手に与えるダメージの指標．固定値．
- 使用可能な術技
  そのキャラクタが選択できる行動の種類．固定．

5.2 行動とその効果

各キャラクタの行動は，術技（行動の種類）と対象キャラクタの組合せで表現される．術技それぞれの効果は以下の通りである．

- 攻撃
  対象キャラクタに攻撃力と同じ値のダメージを与える
- 回復
  精神力を一定量消費して，対象キャラクタの体力を一定量回復する．この術技を使うキャラクタに消費される分の精神力が残っていない場合は選択できない．

5.3 状態遷移

このゲームでは複数のキャラクタが敵チーム・味方チームに分かれて，チーム同士で戦闘を行う．戦闘は戦闘終了条件を満たすまでターンを繰り返し行う．1 つのターンは以下 3 つのステップに分けられる．

(1) **行動選択ステップ**
行動可能な全キャラクタが現在の状態に基づき行動を選択する．全員が行動を選択したら次のステップに進む

(2) **行動処理ステップ**
ランダムな順番で，各キャラクタが選択した行動を順次処理していく．処理順の関係で，あるキャラクタの行動が処理される前にその行動主体のキャラクタ，もしくは行動対象のキャラクタが行動不能になってしまった場合，そのキャラクタの行動は処理されない．全員分の行動を処理したら次のステップに進む．

(3) **戦闘終了判断ステップ**
どちらかのチーム全員が行動不能になった場合に戦闘を終了する．両チームに行動可能なキャラクタが 1 人でもいる場合は(1)のステップへ戻る．

ⓒ 2013 Information Processing Society of Japan

5.4 遷移例

図 3 はこのゲームにおける状態遷移の一例で，左側が遷移前の状態，右側が遷移後の状態を表している．表 2 は図 3 における，状態遷移によるパラメータの変化と，行動選択ステップで各キャラクタが選択した行動の表である．

行動処理ステップでの処理順番が，敵・僧侶・剣士の順番であった場合，最初に敵の攻撃で僧侶の HP は 0 となり行動不能となる．次に僧侶の行動順が来るが，先の攻撃で行動不能となっているので，僧侶の行動は処理されない．最後に剣士の行動が処理され敵にダメージを与えて，図 3 の右側の状態に遷移する．

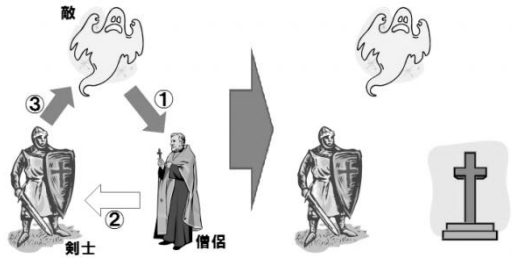

図 3　ゲームにおける状態遷移の例

	剣士	僧侶	敵
HP	50 → 50	20 → 0	60 → 20
MP	50 → 50	80 → 80	10 → 10
攻撃力	40	20	20
選択行動	攻撃	回復	攻撃

表 2　図 3 の状態での各キャラクタのパラメータなど

6. 提案手法

本研究では，人間プレイヤの選択した行動からそのプレイヤの効用を推定し，自身の行動選択に活用するチームメイト AI を構成する手法を提案する．

提案するチームメイト AI の全体像は図 4 の通りである．このチームメイト AI は人間プレイヤの効用を表現するための効用モデルを持っている．学習部分では人間プレイヤの効用を近似するように効用モデルを調整し，行動選択部分では効用モデルを使い自身の行動を決定する．

6.1 効用モデルと期待効用値

効用モデルには，ニューラルネットワークなどに代表される入出力モデルを使用する．この入出力モデルは 4 章で説明した“状態を構成する要素”を入力とし，出力が状態の効用値となる．状態sで行動aを取った場合の期待効用値Euは(1)式で求められる．

$$Eu(s,u,a)=\sum_{s'}\{u(s')Tr(s,s',a)\} \tag{1}$$

ここで，uは状態から効用値を算出する効用関数（効用モデル），$Tr(s,s',a)$は状態sで行動aを取ったとき“未来の状態”s'に遷移する確率である．5 章で説明したゲームのように，同時着手性があるゲームの場合は(1)式に他者の行動bを加えた(2)式で期待効用値を計算することができる．

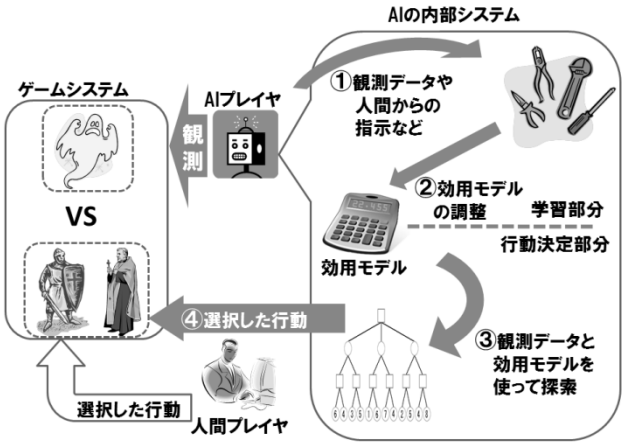

図 4　提案するチームメイト AI の全体像

$$Eu(s,u,a,b)=\sum_{s'}\{u(s')Tr(s,s',a,b)\} \tag{2}$$

6.2 学習部分

学習段階では人間プレイヤの効用を近似するために効用モデルを調整する．学習を行うため事前に学習用のデータとして，効用を近似させたい人間プレイヤの行動データ$d_i=(s_i,a_i)$を複数用意し，これを学習用データの集合$D(\ni d_i)$とする（図 4 の①）．d_iは「人間プレイヤが状態s_iで行動a_iを選択した」という情報である．学習段階は次の 4 つのステップから構成される．

(1) 期待効用値の計算

状態sで人間プレイヤが選択できる行動の集合をAとし，任意の要素を$a_n\in A$とする．同時着手ゲームの場合はさらに，他者が選択できる行動の集合をBとし，任意の要素を$b_m\in B$とする．

同時着手ゲームの場合は(a_n,b_m)の組合せ全てについて(2)式で期待効用値を計算する．この様にして得られた表を期待効用値表$EuTable$とし，(3)式のように表す．

$$EuTable(u,s)=\{Eu(s,u,a_n,b_m)\}_{n,m} \tag{3}$$

表 1 は期待効用値表$EuTable(u,s)$の一例であり，図 2 が$EuTable(u,s)$を計算する際のイメージとなっている．

用意した学習用データに含まれるs_i全てについて$EuTable(u,s_i)$の計算を行い，$EuTable(u,s_i)$の集合である

$$EuTables(u,D)=\{EuTable(u,s_i)\}_i \tag{4}$$

を得る．

(2) 選択確率の計算

前ステップで計算した期待効用値表$EuTable(u,s_i)$から人間プレイヤが状態s_iで行動a_iを選択する確率$p(s_i,u,a_i,EuTable)$を計算する．$p(s_i,u,a_i,EuTable)$の計算方法として様々な方法があるが，気を付けなければな

らないのは，計算方法を「人間らしい」方法にすることである．この算出方法が「人間らしくない」場合，そのチームメイト AI の性能は低いものとなってしまう．本論文でのpの計算の詳細は 7.2.2(2)で示す．

(3) **対数尤度の計算**

前ステップで計算した選択確率から，効用モデル u の対数尤度を計算する．計算式は(5)式の通りである．

$$logLikelihood(u, D) = \sum_i \log\{\varepsilon + p(s_i, u, a_i, EuTable)\} \quad (5)$$

ここで，$\varepsilon(>0)$は$p(s_i, u, a_i, EuTable)$が 0 になってしまう場合を考慮した調整用パラメータである．

(4) **効用モデルの調整**

効用モデルの調整方法としては，ローカルサーチや遺伝的アルゴリズムといった生成検査法，バックプロパゲーションなどの勾配法が挙げられる．2 つの効用モデル u と u'について比べたとき，$logLikelihood(u', D) > logLikelihood(u, D)$ならば，$u'$の方が学習用データの集合$D$を$u$より尤もらしく説明している．つまり，よりサンプルとなった人間の効用に近いということである．このステップでは，対数尤度を最大化するように効用モデルを最適化していく（図 4 の②）．

6.3 行動決定部分

行動決定部分では，自身が持つ効用モデル u_t を使用して行動選択を行う（図 4 の③と④）．ここではまず現在の状態 s_t から期待効用値表$EuTable(u_t, s_t)$を計算し，人間プレイヤがストレスを感じないように行動を選択しなければならない．

$EuTable(u_t, s_t)$から自身の行動を決定する方法として様々な方法があるが，4 章で説明した「人間のストレスとなりうる AI プレイヤの行動選択」については気を付けなければならない．詳細は 7.2.3 節で述べる．

7. 実験

提案手法のチームメイト AI を実装し，仲間プレイヤの効用を近似する実験を行った．効用モデルを近似させる対象として，本論文では人間ではなく戦闘終了時の状態に対して効用を持つモンテカルロ AI を用意した．

提案手法の AI にモンテカルロ AI の効用を学習させるため，提案手法の AI とモンテカルロ AI とでチームを組ませて 50 戦分の戦闘を行い，調整された効用モデルを評価した．

7.1 戦闘参加キャラクタの設定

今回，実験を行うために設定した敵・味方キャラクタのパラメータは表 3 と表 4 の通りである．敵キャラクタは環境の一部に過ぎず，ポイントは味方キャラクタ 1 がキャラクタ 2 の効用を学習して協調できるか，という点である．

	味方キャラクタ 1	味方キャラクタ 2
最大 HP	150	100
最大 MP	50	100
攻撃力	40	20
使用可能術技	攻撃 回復	攻撃 回復
操作 AI	提案手法 AI	モンテカルロ AI

表 3　味方チームに属するキャラクタのパラメータ

	敵キャラクタ 1	敵キャラクタ 2	敵キャラクタ 3
最大 HP	60	60	80
最大 MP	0	0	60
攻撃力	40	40	10
使用可能術技	攻撃	攻撃	攻撃 回復
操作 AI	ランダム AI	ヒューリスティック AI	原始モンテカルロ AI

表 4　敵チームに属するキャラクタのパラメータ

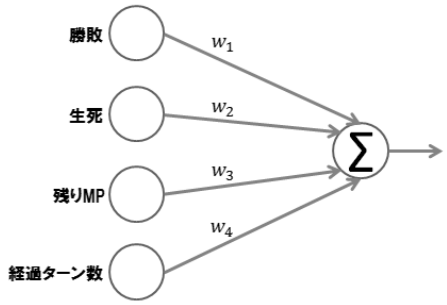

図 5　入出力モデルのイメージ

7.2 提案手法を実装したチームメイト AI の詳細

7.2.1 効用モデル

6 章で説明した提案手法に則り，チームメイト AI を実装した．今回実装したチームメイト AI は“未来の状態”として「戦闘終了時の状態」を評価し，状態の良さを効用値として数値化する AI とした．“状態を構成する要素”として以下 4 つの要素を使用した．

① 戦闘の勝敗

自チームが戦闘に勝利した場合は 1 を，戦闘に敗北した場合は 0 を入力値とする

② 自チームのキャラクタの生死

戦闘終了時，自分を含めた自チームのキャラクタの生死によって入力値が変わる．今回の実験の設定では自チームは 2 人なので，2 人共生きている場合は 1 を，1 人だけ生きている場合は 0.5 を，2 人共行動不能な場合は 0 を入力値とする．

③　チームの残り MP

自チーム全体でみた MP の残りパーセンテージを入力値とする．

④　経過ターン数

入力値は(6)式のシグモイド関数で計算する．

$$Sigmoid(x) = \frac{1}{1 + Exp(-8(x-1))} \quad (6)$$

図 5 は実装した効用モデルのイメージである．効用モデルの調整は，各入力ノードと出力ノードを繋ぐコネクションの結合強度w_jを変化させることで行った．

7.2.2 学習部分

今回の実験では「味方キャラクタ 2」を人間プレイヤに見立てて効用の近似を行うこととする．実験の第一段階で，提案手法の AI は「味方キャラクタ 2」とチームを組み，50 戦分戦う中で学習用データDを収集し学習を行う．図 6 は学習部分全体のイメージである．詳細は以降で説明する．

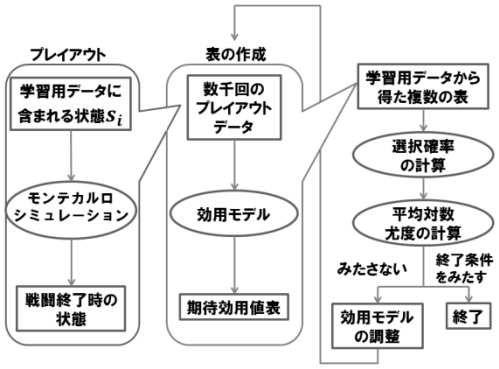

図 6　実装した学習部分全体のイメージ

(1)　期待効用値の計算

「味方キャラクタ 2」が選択可能な行動の集合をA，自身が選択可能な行動の集合をBとして$EuTables(u, D)$を計算する．今回は状態遷移における分岐数の多さ・ランダム性の強さなどの理由から，モンテカルロシミュレーションを使って各行動の組合せ(a_n, b_m)の期待効用値を近似し，期待効用値表$EuTable(u, s_i)$を得る．モンテカルロシミュレーションの処理手順は以下の通りである．

i)　最初のターンの行動選択

状態s_iから戦闘のシミュレーションを開始する．最初のターン，自チームのキャラクタは期待効用値を調べたい行動の組合せ(a_n, b_m)の行動を選択する．敵チームのキャラクタはランダムに行動選択を行う．

ii)　次ターン以降の行動選択

次ターン以降は，戦闘に参加する全キャラクタがランダムに行動を選択する．これを戦闘終了まで繰り返す．

iii)　戦闘が終了したら

戦闘が終了したら効用モデルを使って戦闘終了時の状態を評価し効用値を計算する．

(2)　選択確率の計算

ここでは前段階で得られた期待効用値表$EuTable(u, s_i)$それぞれに対して，人間側の選択肢の選択確率を計算する．今回の実験では「人間らしい」選択確率の計算方法として，以下のような方法で状態s_iにおいて人間プレイヤが行動a_iを選択する確率$p(s_i, u, a_i, EuTable)$を計算した．

i)　人間側の選択肢の削減

効用値にはその人の主観が入っていること，また，人間は完全に合理的ではないことなどから，ゲーム理論などで提唱されている考え方を適用すれば人間の選択確率を正しく計算できるとは限らない．また，今回の AI は期待効用値の計算にモンテカルロシミュレーションを使用しているので，シミュレーションによる誤差も考慮しなければならない．そこで，この段階では「人間らしい」考え方として「明らかな被支配戦略」と「仮の弱被支配戦略」を定義し，これらに該当する選択肢を以降の計算の対象から外すことで，選択確率を計算するための前処理を行う．

「明らかな被支配戦略」を以下の通り定義する．

“相手がどの戦略を取ったとしても，ある戦略の期待効用値の有意水準 5%の下限以下の期待効用値しか得られない戦略”

表 5 は「明らかな被支配戦略」の一例である．各行動の組合せは 100 回ずつモンテカルロシミュレートされているものとする．この場合，(a_1, b_1)の有意水準 5%の下限は 0.4，(a_1, b_2)の有意水準 5%の下限は 0.216 となり，いずれも(a_2, b_1) と(a_2, b_2)の期待効用値より大きくなる．よってa_2は「明らかな被支配戦略」となる．

	b_1	b_2
a_1	0.5	0.3
a_2	0.3	0.0

表 5　「明らかな支配戦略」の一例

「仮の弱被支配戦略」以下のように定義する．

“自分側の 2 つの戦略について見たとき，ある相手の戦略に関して低い方の期待効用値が高い方の期待効用値の有意水準 5%の下限より高い場合，それらを同程度と見なす．このとき同程度と見なした値以外に注目し，相手がどの戦略を取ったとしても，ある戦略の期待効用値の有意水準 5%の下限以下の期待効用値しか得られず，かつ平均期待効用値が 90%以下の戦略”

表 6 は「仮の弱被支配戦略」の一例である．各行動の組合せは 100 回ずつモンテカルロシミュレーションされているものとする．この場合，(a_2, b_1)と(a_1, b_2)共に有意水準 5%の下限は 0.4 となる．相手がb_2の戦略を取った場合，(a_2, b_2)の期待効用値 0.45 は(a_1, b_2)

の有意水準 5%の下限 0.4 より大きいので，これらを同程度とみなす．(a_2, b_1)の有意水準 5%の下限 0.4 は(a_1, b_1)の期待効用値 0.1 より大きい．各戦略の平均期待効用値も 0.3≦0.475×0.9 を満たしている．よってa_1は「仮の弱被支配戦略」として以降の計算の対象から外れる．

	b_1	b_2
a_1	0.1	0.5
a_2	0.5	0.45

表 6　「仮の被支配戦略」の例

以上のようにして削減した人間側の選択肢の集合を$A_{cut}(\subseteq A)$とし，任意の要素を$a_c(\in A_{cut})$と表す．

ii) 人間側の選択肢の評価値計算

前段階で削減されなかった選択肢について，それぞれの評価値を計算する．この評価値は，人間がその選択肢に対して抱く期待度を数値化したものである．計算式は(7)式の通りである．

$$Em(a_c, s_i) = max_c - \left((1 - belief) * (risk_c - average_c)\right) \quad (7)$$

ここで，$\max_c$はa_cを選んだ場合に得られる可能性がある最大期待効用値，$average_c$はa_cを選んだ場合の平均期待効用値，$risk_c$はa_cを選んだ場合の最大期待効用値と最低期待効用値の差の絶対値，$belief$は人間プレイヤから AI プレイヤに対する信頼度を表す調整用パラメータである．本研究では，ある程度の期間共にプレイしてきた人間プレイヤと AI プレイヤを想定しているので，$belief$の値を 0.5 とした．$belief$値が高ければ最大値，低ければリスク回避的な選択を行う．

iii) 人間側の選択肢の確率を計算

状態s_iにおいて人間プレイヤが行動a_iを選択する確率を(9)式で計算する．Em値が 1 違うと選択確率が e 倍になるような式である．

$$sumEm = \sum_c Exp\left(Em(a_c, s_i)\right) \quad (8)$$

$$p(a_i, s_i) = \begin{cases} \dfrac{Exp\left(Em(a_i, s_i)\right)}{sumEm} & (\exists a_i \in A_{cut}) \\ 0 & (otherwise) \end{cases} \quad (9)$$

(3) **対数尤度の計算**

対数尤度は，εを 0.1 として(5)式の通りに計算する．

(4) **効用モデルの調整**

今回はローカルサーチにより効用モデルを調整した．現在の効用モデルをu_t（図 5 を参照）とする．

ローカルサーチでは最初に，u_tの入力ノードと出力ノードを繋ぐコネクションの中からランダムにひとつのコネクションを選択肢し，その結合強度を 0.8～1.2 倍の範囲でランダムに変化させ，その後最大強度が 1 になるよう正規化する．

次に，全ての結合強度から絶対値が最大となるもの（$maxAbs$）に注目し，この絶対値を使い全ての結合強度を(10)式のように正規化する．w_i'は正規化後の各結合強度である．

$$w_i' = w_i \div maxAbs \quad (10)$$

上記のようにして，より尤度の高い効用モデルの候補u'を生成する．u'を使用して学習用データ集合Dの対数尤度を計算し，この値が現在の効用モデルu_tの対数尤度よりも大きかった場合，u_tをu'と置き換える．

7.2.3 行動選択部分

現在の状態をs_t，現在の効用モデルをu_tとする．行動選択部分では，最初にこの AI が持つu_tを使い，人間プレイヤが選択可能な行動と，自身（「味方キャラクタ 1」）が選択可能な行動の組合せについての期待効用値表$EuTable(u_t, s_t)$を求める．このとき，7.2.2 で紹介したモンテカルロシミュレーションを使い，期待効用値の近似を求めている．$EuTable(u_t, s_t)$が求まったら，最大期待効用値となった組合せに含まれている自身の行動を，このターンの行動として選択する．

7.3 その他の AI の概要

提案手法以外の，実験で使用する AI の概要を説明する．

(1) **モンテカルロ AI（MCAI）**

モンテカルロ AI には戦闘結果について簡単な効用を持っており，自身と味方（この実験の場合は「味方キャラクタ 1」）が選択できる行動の組合せについてモンテカルロシミュレーションを行う．得られた期待効用値表から期待効用値が最大となる行動の組合せに含まれる自身の行動を選択する．モンテカルロ AI が持つ効用の内容は次の通りである．

「戦闘終了時，自チームのキャラクタ 2 人中 1 人でも生きていれば 1 の効用を持ち，2 人共行動不能の場合は 0 の効用を持つ．」

(2) **ランダム AI**

ランダム AI はその時選択可能な行動の中からランダムにひとつの行動を選択する AI である．

(3) **ヒューリスティック AI**

このヒューリスティック AI は，敵チームの中で一番 HP の数値が低い敵を狙って攻撃する AI である．

(4) **原始モンテカルロ AI**

原始モンテカルロ AI は自身の行動選択肢のみについてモンテカルロシミュレーションを行い，一番期待勝率が高くなった行動を選択する AI である．

ⓒ 2013 Information Processing Society of Japan

7.4 実験の設定

提案手法の AI の設定は以下の通りである．

- 一回の行動選択で 5000 回（原始モンテカルロの場合は 500 回）のシミュレーションを行う
- シミュレーションの対象となる行動の組合せは UCB 値を使って決定する
- 効用モデルを使用し，シミュレートした戦闘終了時の状態から効用値を算出する
- 期待効用値表中，一番期待効用値が高い組合せに含まれる自身の手を選択する

7.5 実験結果

7.5.1 学習後の効用モデルのパラメータ

提案手法の AI にはモンテカルロ AI の効用を 50 回の戦闘で学習させた．学習終了時の，提案手法の AI が調整した効用モデルの結合強度パラメータは表 7 のようになった．モンテカルロ AI は戦闘の勝敗のみに効用を持っているので，その効用を上手く学習できていると言える．

	効用モデルの結合強度
戦闘の勝敗	1
キャラクタの生死	-6.6×10^{-14}
チームの残り MP	4.8×10^{-15}
経過ターン数	-1.2×10^{-5}

表 7　学習後の結合強度のパラメータ

8. まとめ

本研究では人間プレイヤの効用を推定・学習し，その効用を使って行動選択を行うことで，人間プレイヤの満足度が高い AI プレイヤの構成手法を提案し，独自に設定したコマンド形式 RPG ゲームに適用した．提案手法の AI はシステムの内部に人間の効用を近似するための入出力モデルを持ち，学習部分ではそのパラメータを調整することで人間プレイヤの効用を近似し，行動選択部分では調整した入出力モデルを使うことで，人間が不満に思わないような行動を選択するようにした．

今後の研究予定として，提案手法は実際に人間の不満度を改善できるのかを調べるために被験者実験を行うことや，効用モデルを多層パーセプトロンなどの複雑なモデルに変えて性能を調査することなどを考えている．

参考文献

1) Sander Bakkes, Pieter Spronck and Eric Postma: TEAM:The Team-oriented Evolutionary Adaptability Mechanism, Entertainment Computing - ICEC 2004 (2004), pp.273-282 (2004).
2) Aswin Thomas Abraham and Kevin McGee: AI for Dynamic Team-mate Adaptation in Games, Computational Intelligence and Games (CIG), 2010 IEEE Symposium on, pp.419-426 (2010).
3) Kenneth O.Stanley, Bobby D. Bryant, Student Member, IEEE, and Risto Mikkulainen : Real-Time Neuroevolution in the NERO Video Game, Evolutionary Computation, IEEE Transactions on, Vol.9, No.6, pp.653-668 (2005).
4) 田村坦之，中村豊，藤田眞一: 効用分析の数理と応用，コロナ社, pp.1-28 (1997).
5) 加藤英明: 行動ファイナンス－理論と実証－，朝倉書店, pp.61-66 (2003).