Vol.2021-HCI-192 No.7
2021/3/15

強化学習環境を模した VR 迷路の探索課題における人の学習過程に関する考察

寺井輝 [1]　田辺弘子 [2]　小宮山摂 [2]

概要：迷路を用いた研究は，動物を用いた空間認知プロセスや学習を評価する実験から，人とよく似た学習過程を持つと言われる強化学習を用いたアルゴリズムの研究まで，広く扱われるテーマである．その一方で物理的な制約に伴い，人に対する迷路を用いた研究は手つかずの状態であり，知見として既にある動物や強化学習の迷路学習過程と比べ，どのような違いがあるのかは明らかになっていない．そこで本研究では人の学習過程と強化学習を比較することを目的に，VR 技術を利用し，強化学習においてエージェントがおかれた状況に近い条件で人が迷路探索を行うことができるシステムを作成した．このシステムを用いた迷路探索実験を行い，さらに強化学習手法を用いたシミュレーションと比較を行った結果，人は強化学習のような手順を覚える学習法と，空間情報を活用した学習を組み合わせて迷路学習を行っていると考えられ，学習率を高くした 3 手更新の Multi-step Learning が人の迷路学習時の挙動と比較的近いことが明らかになった．

キーワード：迷路探索，VR，強化学習，認知地図

1. はじめに

学習や記憶は，動物の生命維持や進化にとって重要な脳の機能であり，既存の情報を元にして問題解決や意思決定を行うプロセスに基づいている．マウスなどの小動物を用いた実験では，これまで様々な手法によって学習や記憶が評価されているが，中でも Y 字型迷路試験や Morris 水迷路試験といった迷路探索課題により，それぞれ短期記憶の評価および潜在学習能力の評価を行うことができる[1-2].

人において学習や記憶を評価することは，記憶障害などの病態の解明や運動学習メカニズムの解明に寄与すると考えられるが，現実世界において人が迷路探索課題を行うことは物理的に困難なため，これまで人を対象とした迷路学習や記憶の評価は手付かずであった．

しかし，VR 環境で迷路を作成すれば，先述の物理的制約を回避することが可能であり，人においても学習や記憶評価のための迷路探索課題を行うことができる．

さらに近年では，機械学習の一種である強化学習（reinforcement learning）を用いて迷路を探索，学習させるといった研究が注目を集めている．強化学習は試行錯誤を繰り返して学習を進める様子が人の学習と非常に似ていると話題であり，しばしば両者は比較されることもある．

そこで本研究では VR （Virtual Reality） 技術を用いて，仮想空間内で人が迷路探索を行うことができるシステムを作成し，かつその環境を強化学習に近いものに設定した．以上より，強化学習環境に近い，情報が制限された迷路における人の迷路学習過程について，動物迷路実験の知見や強化学習の枠組みを絡めた検証と考察を行うことを本研究の目的とした．

2. 関連技術・研究

2.1 強化学習と脳科学との関連

Skinner は Skinner box と呼ばれるネズミの行動実験を行い，ある行動により好ましい結果がもたらされると，同じ状況においてその行動を選択しやすくなることを明らかにした[3]．このような生物の行動は，試行錯誤を繰り返して学習を行う強化学習アルゴリズムに非常に近い学習であると言える．また，現在では生物の脳内において強化学習と似た学習モデルを形成しているという説が有力である[4]. Shultz らはサルを対象としたオペラント条件付け実験により，ドーパミンにより脳内で TD 誤差の情報が伝播されていることを提唱した[5].

ドーパミン作動性ニューロンは，大脳基底核の入力部にあたる線条体を主要な投射先にしているため，線条体の神経活動が価値関数の推定に関与していると考えられる: Shidara ら[6]は腹側線条体ニューロンを，Samejima ら[7]は背側線条体ニューロンの活動をそれぞれ記録し，腹側線条体ニューロンの活動は状態価値関数 V の挙動に，背側線条体ニューロンの活動は行動価値関数 Q の挙動に高い相関を持つことを示した．これらの結果を受けて，Barto らは，大脳基底核回路において報酬の TD 誤差による強化学習が行われているというモデルを提案した[8].

一方で，大脳基底核が強化学習法と同様の情報処理を行っていると仮定する場合，学習を上手く進めるためのパラメータ調整を司る器官が存在することが想定されるが，Doya はこれらのメタ学習に神経修飾物質が作用している仮説を提案している[9]．これによると，シナプス可塑性を調節するアセチルコリン[10]が，記憶の保存と更新の間のバランスを制御する学習率 α の役割に相当し，レベルの高さが衝動的な行動の抑制に関連しているセロトニン[11-12]

1 青山学院大学大学院理工学研究科
Graduate School of Science and Engineering, Aoyama Gakuin University
2 青山学院大学　Aoyama Gakuin University

が，報酬の短期的予測と長期的予測のバランスを制御する割引率 γ に相当すると提唱した．

2.2 迷路探索に関する研究

一般に，強化学習を用いて学習した迷路探索の学習モデルをそのまま再利用して，微小量形状を変化させた迷路の学習に適用しても上手く学習が進まないことが分かっている．齋藤らは形状変化させた迷路へ学習結果を再利用させる際に，ある一定の閾値以上の手数を要した際には行動価値関数をリセットするといった手段を用いれば，限定的な環境変化においては効率的に学習結果を再利用ができるとした[13]. また我々は, 環境変化のある迷路において一度学習した学習モデルを再利用する際に，一定値を用いて行動価値関数を底上げすることにより，変更後の迷路において最初から学習をやり直す場合よりも効率的に学習を収束させられることを示した[14].

一方で，生物が迷路探索を行う場合について, Tolman はネズミを用いた迷路学習実験を通じて，迷路学習の際に生物は頭の中で空間情報を地図化しており，認知地図を作り上げていると提唱した[15]. また, 北濱らは人を対象として実際に立体迷路を作成し，迷路探索時の注視行動に着目をした．結果として頭の中で完成された迷路全体の地図を参照しながらゴールに辿り着いたのではなく，注視が何度も行われた風景から感覚的に経路を判断していると提唱した[16].

2.3 強化学習の枠組みを用いた人との比較

佐々木らは強化学習型のタスクを人に対して行わせ，強化学習アルゴリズムの枠組みから人間の行動決定を明らかにしようと試みた[17]. 結果, マクロ的には人の行動決定を強化学習の枠組みで説明できる部分もあったが，人の記憶メカニズムなどの要因が学習結果に影響を及ぼし強化学習の枠組みだけでは捉え切れないものであると結論付けた．

また，小川らは経済実験を例に，強化学習がどの程度被験者行動を説明するかを検討した[18]. 結果, 強化学習モデルのみでは上手く説明できない行動も存在し，強化学習以外の意思決定アルゴリズムも構想する必要があると唱えた．

3. 強化学習手法について

3.1 Q 学習（Q-learning）

Q 学習は TD 法（Temporal Difference Learning）の一種であり，現時点で予測される報酬を基に 1 回の行動直後に Q 値を更新する手法である．また，Q 値の更新が方策に依存しないのが特徴であり，これにより，学習の収束を早める効果が見込まれる．時刻 t において，状態 s_t における行動 a_t の価値 $\mathrm{Q}(s_t, a_t)$を Q 学習によって更新すると以下の式（1）の通りとなる．

$$\mathrm{Q}(s_t, a_t) \leftarrow \mathrm{Q}(s_t, a_t) + \alpha * (r_{t+1} + \gamma \max_a \mathrm{Q}(s_{t+1}, a) - \mathrm{Q}(s_t, a_t)) \quad (1)$$

ここで，α は学習率$(0 < \alpha \leq 1)$であり，今行動した結果を Q 値にどれほど反映させるかを調整するパラメータである．一般に α の値が小さいほど学習速度は遅いが学習は安定することが分かっている．

3.2 モンテカルロ法（Monte Carlo Methods）

モンテカルロ法は 1 エピソードが終了してから，実際に得られた報酬を基に一気に Q 値を更新する手法である．更新式は以下の式（2）となる．

$$\mathrm{Q}(s_t, a_t) \leftarrow \mathrm{Q}(s_t, a_t) + \alpha * \left((r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{T-1} r_T) - \mathrm{Q}(s_t, a_t)\right) \quad (2)$$

ただし，時刻 T はエピソード終了時刻である．

3.3 Multi-step Learning

Q 学習が 1 手ごとの Q 値更新, モンテカルロ法が 1 エピソード終了時に Q 値更新であったが，その中間のタイミングで Q 値を更新することも可能である．この手法を Multi-step Learning あるいは n-step learning といい，n 手ごとに Q 値を更新する場合は以下の式（3）となる[19-21].

$$\mathrm{Q}(s_t, a_t) \leftarrow \mathrm{Q}(s_t, a_t) + \alpha * \left(\left(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^n \max_a \mathrm{Q}(s_{t+n}, a)\right) - \mathrm{Q}(s_t, a_t)\right) \quad (3)$$

3.4 softmax 方策

softmax 方策は Q 値に softmax 関数を適用することにより, Q 値の高い順に行動選択確率が与えられる方策であり，状態 s における行動 a の選択確率 P を以下の式（4）で決定するものである．

$$P(s, a) = \frac{exp\,(Q(s, a)/\tau)}{\sum_{a' \in \mathcal{A}(s)} exp\,(Q(s, a')/\tau)} \quad (4)$$

ただし，τ は温度パラメータ（$\tau > 0$），$\mathcal{A}(s)$ は状態 s における選択可能な行動の集合を表す．学習が進むにつれて温度パラメータ τ を下げることにより, 行動価値の低い行動を抑制することができる．

4. VR 迷路実験

4.1 VR 迷路システム

Unity を使用して仮想環境内に迷路を作成した．迷路は外壁および内壁のみで構成すると先が見通せるため，強化学習の環境と大きく異なってしまう．そこで一辺 6m からなる立方体の部屋とそれらを繋ぐ通路によって構成し，強

化学習における学習状況に近づけるために隣の部屋の様子が見えない設計とした．

HMD（Head Mounted Display）にはHTC Viveを使用し，利用者の首振りに合わせてカメラが連動して動く設計となっている．迷路内の移動にはHTC Vive付属のコントローラを用い，視界前方に隣の部屋への通路がある状態でコントローラのトリガーボタンを押下すると隣の部屋まで自動で前進を行う．これにより，実際の迷路内を歩く必要がなく，その場で回転をするのみで没入感のある迷路探索を体験する環境を作成した．システムを実際に使用している様子が以下の図1である．

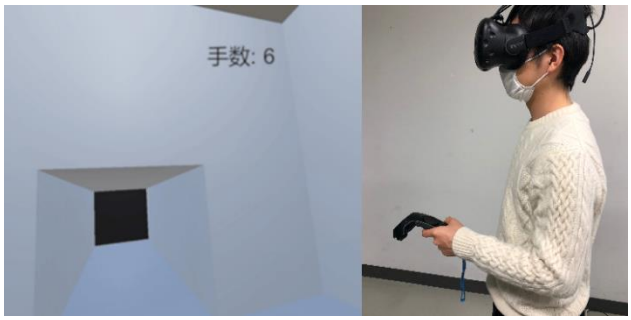

図 1: VR 迷路システム利用時の様子

4.2 作成した迷路について

6×6サイズ，全36部屋からなる迷路を，壁伸ばし法アルゴリズムを用いて無作為に2種類作成した（以下，それぞれ迷路1，迷路2とする）．壁伸ばし法は単純な迷路生成アルゴリズムとしてよく使われる手法であり，迷路内の格子点からランダムに1点を選び既存の壁にぶつかるまでランダム方向に壁を伸ばし続けることで迷路を作成する．これを繰り返し，すべての格子点について壁生成を行うまで続ける．これにより，閉区間がなく，特定の2点を繋ぐ最短経路が1種類に限定された迷路が出来上がる．

作成した迷路の概要図を図2に，VR迷路の平面図を図3に示す．また，それぞれの迷路について無作為に3か所壁を選択し，変更後迷路で通行可能とした．これは，一度学習した結果を用いた効率的な学習が可能かを検証する狙いがある．迷路1,2ともに最短経路は初期状態では14手，変更後迷路では10手となる．概要図の点線部と平面図の水色の通路が後に通行可能となる壁である．

図 2: 迷路 1（左）と迷路 2（右）

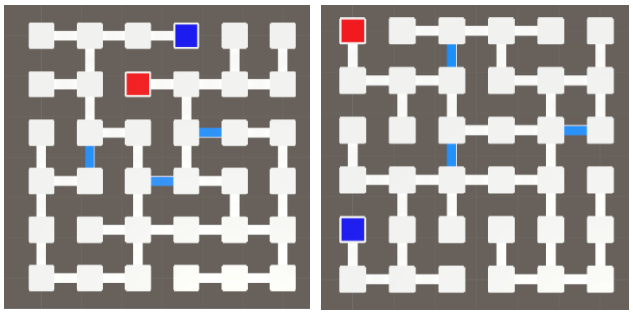

図 3: 実際に探索する迷路 1（左）と迷路 2（右）

4.3 迷路探索タスク

迷路のスタート地点からゴール地点までの最短経路を学習するタスクを設定した．ただし，1度の探索（以下，これを1エピソードとする）における最大移動数は最短経路の手数である14手の約1.5倍である20手に設定した．20手に到達した場合はスタート地点に戻されて再度学習を始める．同様に20手以内にゴールに到達した場合にも再度スタート地点より学習を始める．探索を繰り返す中で被験者が最短経路であると確信した上で，その経路を2度続けて辿ることができた時点で学習が完了したとみなしてタスクを終了とする．なお，被験者にはあらかじめ最短経路の手数は教えておらず，学習回数に上限は設けなかった．

4.4 実験方法

20代の大学生10名を用いて，各迷路5名ずつ迷路探索を行った．また，実験中の被験者の体調には十分に配慮し，被験者が体調不良を訴えた場合や攻略不可と判断した場合にはその場で直ちに実験中止とした．

学習が終了後，迷路の形状を微小量変更させ，初回の学習結果を活かすよう指示をした上で再度迷路探索タスクを行わせた．取得するデータは学習回数，行動履歴および実験終了後に行うアンケートである．

4.5 実験結果

4.5.1 学習回数一覧

各被験者が各迷路で行った学習回数の一覧は以下の表1の通りである．被験者1-5が迷路1，被験者6-10が迷路2で実験を行った．概ねどの被験者も変更後迷路における学習回数は少ないという結果が得られたが，迷路1の被験者3,4が変更後迷路で更新された最短経路を学習することができなかった．また，迷路1においては被験者5が体調不良により初期迷路における学習完了時点で実験を中止，迷路2においては被験者10が攻略不可により途中で実験を中止した．

表 1: 各被験者の学習回数一覧

	被験者	初期迷路 学習回数	変更後迷路 学習回数	変更後迷路 学習可否
迷路 1	1	19	4	○
	2	12	6	○
	3	11	2	×
	4	6	3	×
	5	10	-	×
迷路 2	6	9	4	○
	7	9	4	○
	8	12	5	○
	9	12	19	○
	10	19（中止）	-	×

4.5.2 各部屋通過回数の一覧

次に初期迷路と変更後迷路について，被験者が各部屋を通過した回数をヒートマップ表示した．変更後迷路で最短経路を学習することができた被験者の中から各迷路 1 名ずつ被験者 1，6 と，学習ができなかった被験者 3，4 をピックアップしてそれぞれ図 4，5 に示す．ただし，S，G の表記がそれぞれスタート地点とゴール地点を示し，各部屋の数値はその部屋を通過した回数を示す．

被験者 1 と 6 はショートカットとなる通路を通過できているのに対して，被験者 3 と 4 は変更後迷路においても初期迷路で覚えた道順をそのまま辿っているのが分かる．

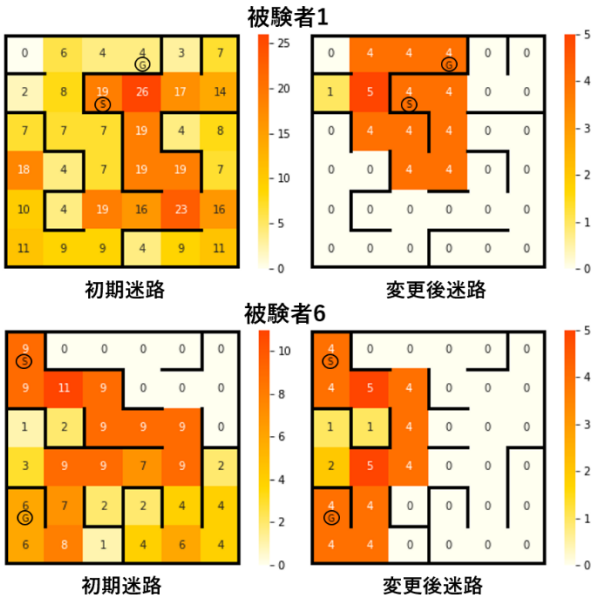

図 4: 被験者 1，6 の各部屋通過回数ヒートマップ

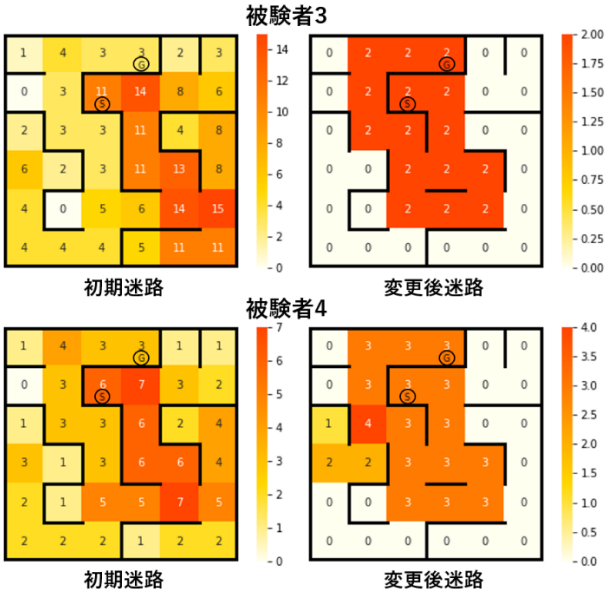

図 5: 被験者 3，4 の各部屋通過回数ヒートマップ

4.5.3 経路図

被験者 1 の実際に辿った経路を図示したものが以下の図 6 である．各エピソード内で辿った経路を赤矢印で示した．

被験者 1 を含む各被験者に共通して見られた傾向として，初期迷路においてゴールに初めて到達するまでは各エピソード間で大きく経路を変更するのに対し，一度ゴールに到達してからは経路変更が小さくなっていた．

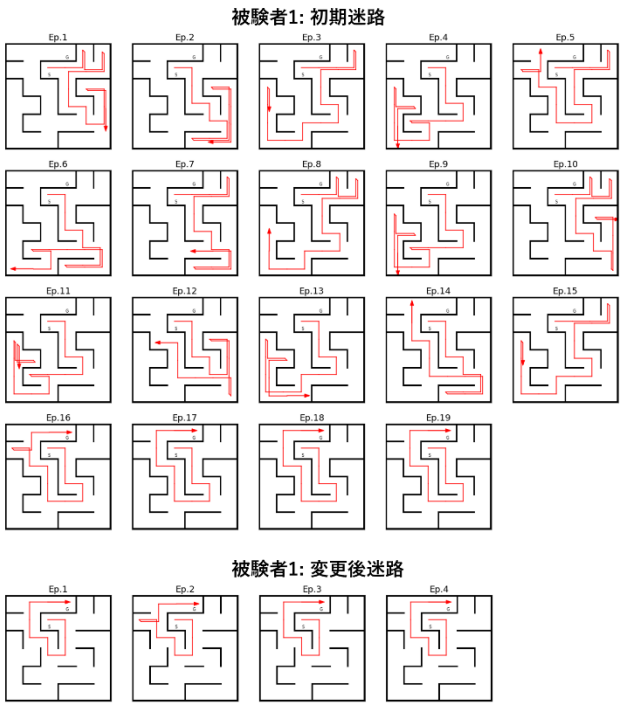

図 6: 被験者 1 の経路図

4.5.4 認知地図の描画結果

実験後に各被験者に認知地図が描けているかの確認として，学習したスタート地点とゴール地点の位置関係を描画させた結果が以下の図 7 である．ただし数値は被験者番号を示す．

迷路 2 は全被験者がゴール地点の方角はおおむね正確に認識できていたのに対し，迷路 1 はゴール地点を正確に把握できた人はいなかった結果となった．また，変更後迷路で最短経路を学習することができなかった被験者 3 と 4 のゴール地点の認識も大きく異なっていた．

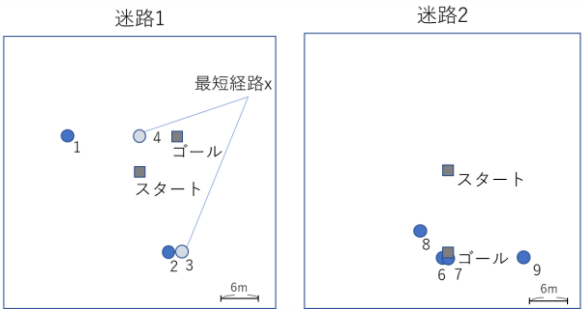

図 7: 実験後の各被験者が描いた認知地図

4.5.5 アンケートに関する結果

実験後に自由記述形式のアンケートを行い，主に探索時の主観的な攻略法を調査した．その中で複数の被験者から共通して挙げられた内容が以下の 3 点である．

・「なるべく同じ道は通らないようにした」 被験者: 4, 6, 7
・「明らかに行き止まりだった部分には再度行かないようにした」 被験者: 3, 4, 5, 7
・「手順を覚えるように攻略をした」 被験者: 2, 3, 4, 5, 8, 9

5. 強化学習による各迷路のシミュレーション

5.1 シミュレーション内容

第 4 章で作成した迷路 1 と 2 の初期迷路に対して強化学習を用いたシミュレーションを行い，作成された学習モデルや学習時の挙動を基に，人の学習時の挙動と強化学習の学習時の挙動の比較を行っていく．

強化学習手法は Q 学習，モンテカルロ法，Multi-step Learning の 3 種類を使用した．学習内容は第 4 章の迷路探索タスクと同じであり，スタート地点からゴール地点までの最短経路を最大 20 手以内に探し出すものとした．また，学習終了の条件は最短経路でゴールに到達したエピソードが 4 エピソード連続して発生した場合とした．報酬の設定としては，

1. ゴール時に正の報酬 +1.0
2. 非ゴール時，一手ごとに微量の負の報酬 -0.1
3. 行き止まり時に負の報酬 -0.3

の 3 つを設定した．これは最短経路を探し出すタスクであること，および第 4.5.5 項のアンケート結果より，「一度行き止まりだと判明した部分にはなるべく入ろうとしなかった」とあったことから一手ごとと行き止まりの際に負の報酬を設定した．また，初期の Q 値は全て 0 で初期化をした．

5.2 各手法間における学習回数の変化

迷路 1 と 2 を Q 学習，モンテカルロ法，Multi-step Learning を用いて各 50 回ずつ迷路探索タスクを行った．Multi-step Learning の更新タイミングには 2 手から 19 手までを使用した．また，各種パラメータは一般によく用いられる値の割引率 $\gamma = 0.9$，学習率 $\alpha = 0.1$を設定してある．

各迷路における学習回数と，初ゴールから学習完了に要する回数を一覧表示したものが図 8 である．ただし，人の実験においては学習終了の条件が最短経路を 2 度続けて辿ることができた時と設定したため，本実験においても最短経路が 4 連続したうちの 2 エピソード目を学習終了時として調整をしてある．

学習回数について，一元配置分散分析を行ったところ主効果が見られ，多重比較の結果，どちらの迷路においても Q 学習を用いた際の学習回数は，Q 学習以外の各手法におけるものよりも有意に多い（$p < 0.001$）という結果となった．Multi-step Learning については更新間隔が 3 から 5 手の間に最小学習回数を取ることが多く，それ以上の間隔では，更新の間隔が長くなるにつれて学習回数が多くなる傾向が見られた．

次に初ゴールから学習完了までにかかる回数について，一元配置分散分析を行ったところ主効果が見られ，多重比較の結果，Q 学習を用いた際の初ゴールから学習完了までにかかる回数は，Q 学習以外の各手法におけるものよりも有意に多い（$p < 0.001$）という結果となった．また，更新タイミングが 5 手あたりから初ゴールから学習完了までにかかる回数がほとんど一桁回数といった数値となり，一度ゴールに到達するとその後は余計な挙動をせず一気に最短経路をなぞるような最適化をしていることが伺える．

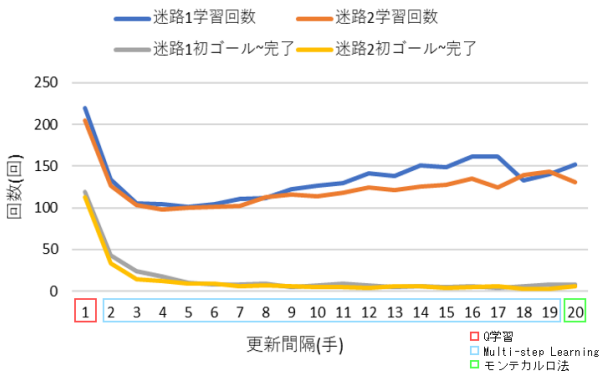

図 8: 各迷路における手法別の平均学習回数一覧と初ゴールから学習完了までにかかる平均回数一覧

5.3 学習率を変更した際の学習回数の変化

次に学習率 α を変更した際の学習回数の変化を観察した．用いる手法は Q 学習と第 5.2 節で平均学習回数が最も少なかった区間より 3 手更新の Multi-step Learning の 2 手法である．更新間隔の長いモンテカルロ法については学習率を上げた際に上手く学習が進まなくなり，学習が収束しなかったため除外した．それぞれの手法について 50 回ずつ迷路探索タスクを行い，平均学習回数を検証する．

各迷路における上記 2 手法の学習率別平均学習回数をそれぞれ示したものが図 9 である．どちらの手法も学習率を上げることにより学習に必要な試行回数を大幅に減少することができている．また， 3 手更新の Multi-step Learning の方が Q 学習と比較して学習回数が少なく済んでいることも特徴的である．

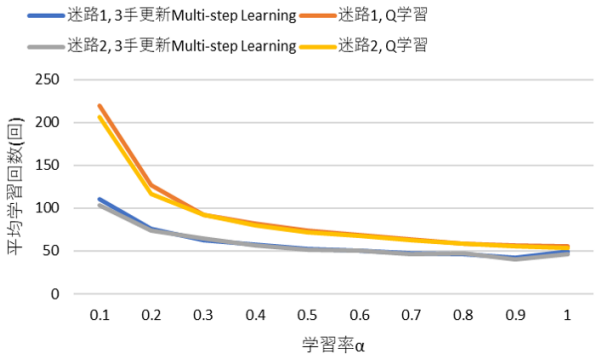

図 9: 迷路 1，学習率の違いによる学習回数の変化

5.4 人の学習結果との比較

第 5.3 節では, Q 学習と 3 手更新の Multi-step Learning の学習率を 0.9 程度まで上げると，初期設定と比べ非常に少ない学習回数で学習完了となった. そこで, 学習率を 0.9 に設定した 2 手法と, 実際の人の迷路探索の結果を比較した. 学習回数と初ゴールから学習完了までの学習回数の比較を行ったものを図 10 に示す．なお，迷路 1 と迷路 2 で同様の結果が得られたため迷路 1 の結果のみを示す.

一元配置分散分析を行ったところ主効果が見られ，多重比較の結果, 人の学習回数は Q 学習と 3 手更新の Multi-step Learning の学習回数よりも有意に少ない（$p < 0.001$）という結果となった.

また，初ゴールから学習完了までにかかる回数についても同様の分析より，Q 学習を用いた場合が，人と 3 手更新の Multi-step Learning の初ゴールから学習完了までにかかる回数よりも有意に多い（$p < 0.001$）結果となった.

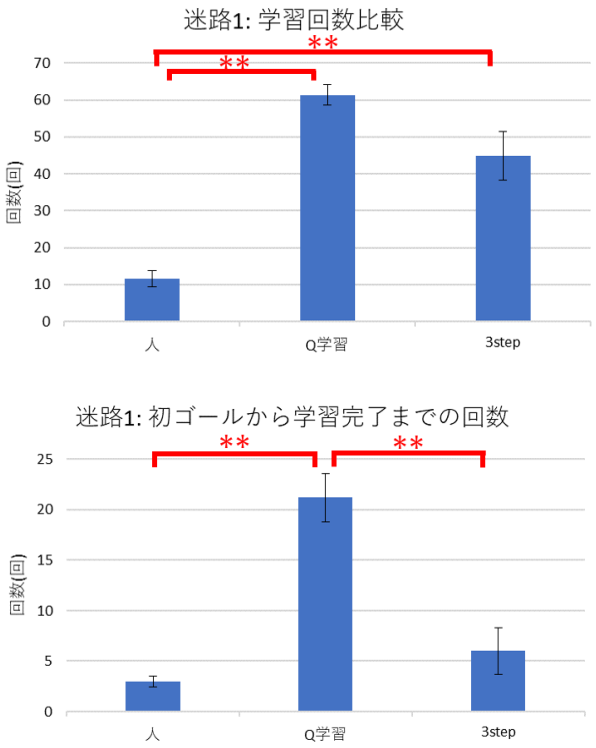

図 10: 人と各学習手法の学習回数比較（上）と
初ゴールから学習完了までの回数比較（下）

5.5 強化学習の経路図

第 5.4 節において，Q 学習と比べて初ゴールから学習完了までにかかる回数が有意に少なく、全体の学習回数も少ない傾向が見られる，3 手更新の Multi-step Learning，学習率 0.9 の設定でシミュレーションした際の初ゴール前後 5 エピソードの経路を図 11 に示す.

人と同様に初めてゴールするまでは経路を大きく変更しているが，それ以降は小さく経路変更をするのみで，少ない回数で最短経路学習が完了しているのが見て取れる．また，同じ経路を行ったり来たりする点が人の挙動とは異なる特徴である.

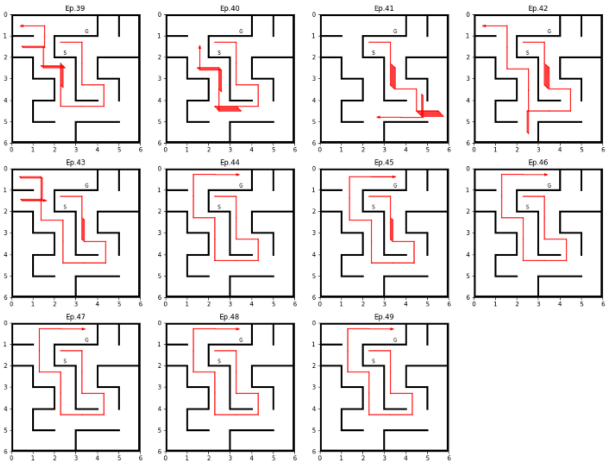

図 11: 3 手更新，学習率 0.9 設定の強化学習
初ゴール前後 5 エピソードの経路図

6. 考察

6.1 人の迷路学習過程についての考察

第 4.5.1 項と第 4.5.2 項の結果より，大半の被験者が初期迷路の学習結果を上手く活用し，変更後迷路においてショートカットを通過する最短経路学習を，非常に少ない試行回数で行うことができると分かった．このことから Tolman の提唱した認知地図の特性[15]のような空間的な情報を活用した学習を行っていることが考えられる．

しかしその一方で，変更後迷路で最短経路を学習できなかった被験者がいることも分かった． 第 4.5.4 項の結果より，被験者 3 についてはゴール地点の認識が大きくずれていた点から，上手く空間情報の活用ができなかったことが想像されるが，被験者 4 は非常にゴールに近い位置を認識することができていた．また，上手く最短経路を学習できた被験者 2 も大きくゴール位置の認識がずれていたことから，認知地図の形成と変更された最短経路の学習可否はあまり関係のない項目であることが推察される．また，第 4.5.5 項の結果を組み合わせてみると，被験者 3,4 はともに「手順を覚えるように攻略をした」という攻略法を取っていたことが分かるが，これは強化学習の学習法に近いと言える．強化学習の学習モデルは何かしらの工夫をしない限り，形状変化させた迷路に一度学習させたモデルを再利用しても上手く学習が進まない[13-14]との知見があるが，まさにその状態と一致している．しかし，同じく手順を覚えるように攻略をした被験者 2 や 8 は変更後迷路においても上手く学習が進んでいることから，必ずしも「手順を覚える攻略」が強化学習のものと完全に同じであるとは言えないだろう．また，もし人の迷路学習過程が空間情報の活用のみによって決まるのであれば，ゴール地点を正確に認識できていない被験者が学習を行うのは困難であると推測される．よって本研究で作成した強化学習環境に近い迷路における探索においては，人は「手順，あるいは動作を覚える」といった強化学習に近い学習法と,「頭の中で位置関係を描くような空間情報の活用」を組み合わせた学習を行っているのではないかと考えた．そして，両者を活用するトレードオフが人により異なり，強化学習型の学習法を強く活用した場合には被験者 3 や 4 のように変更後迷路において更新された最短経路に気付くことができなかったのではないかと考えられる．

また，両者の活用割合を決めるものとして，迷路の大きさと本実験システムの影響も関係があると考えられる．今回の実験では難易度を考慮した 6×6 サイズのものとしたが，もしこのサイズがさらに大きくなった場合，頭の中で認知地図を描くことは困難になり，強化学習型の学習に頼る割合が増加すると思われる．さらに，本実験システムは操作方法がコントローラを用いた移動ということもあり，移動した距離感を掴みにくいことが強く影響してしまった可能性が考えられる．例えば，コントローラによる移動ではなく，実際に足を動かすことにより移動するようなシステムを作成した上で同様の実験を行うことは，今後検証すべき内容である．

6.2 強化学習と比較した人の迷路学習過程の考察

第 5 章のシミュレーション結果より，Q 値の更新間隔を適切に設定し，学習率を高めた場合に学習回数を大きく減少させることが分かった．また，学習率を 0.9 程度まで高めた 3 手更新の Multi-step Learning においては，初めてゴールをしてから学習完了までに要する学習回数が人と有意差の無い結果となった．このことから，人の学習挙動に比較的近いのは上記の手法であると判断した．この結果について人の行動に照らし合わせて考察を行う．

まず，Q 値の更新間隔であるが，第 4.5.2 項より人はゴールに一度到達してからは小さな経路変更を繰り返して最短経路を探るような傾向があることが分かっている．このことから，人は初めてゴールに到達した瞬間に，ゴールまでの大体の経路を頭に思い浮かべることができていると想像できる．つまり，ゴール時に得られる報酬がそのエピソード終了時に複数の状態に伝播している状態であると考えられる．よって，ゴールの報酬を 1 エピソードに 1 状態までしか伝えることのできない Q 学習は人の学習過程とはやや異なると言えるだろう．一方で，モンテカルロ法や，更新間隔の長い Multi-step Learning では，1 エピソードで多くの状態に報酬を伝播させることが可能であるが，更新が行われるまではそれまでの探索結果を活かすことができない．例えば行き止まりに入るなどの明らかにマイナスな行動でも，設定した更新間隔になるまでは学習することができない．これは人の学習挙動から見ても不自然である．これらを踏まえて更新間隔が 3 手である場合を考えると，複数状態へ報酬を伝播させることが可能であり，例えば袋小路に入ってしまった際には袋小路入口の辺りから，この道は間違っているという情報を学習することもできる．この感覚は実際に人が迷路探索を行う場合にも妥当な学習感覚であると考えられる．

次に学習率について見ていく．学習率を上げることは毎回の学習結果を強く反映させるということになる．これを人の迷路探索時の行動に当てはめると，直前に通った経路の記憶が一番鮮明残っており,「さっきこの道は行き止まりだったからここは絶対に違う」といった考え方に相当するだろう. これにより, 直前に通った行き止まりの道は避け, 経路を大きく変更するが，しばらく経つと行き止まりの記憶が薄れてまた同じ道に入ってしまう，といった行動が起こりうると想像できる．さらに，学習率が大きいことによりゴール時の報酬を大きく伝えることにも役立つ．また，第 5.5 節の結果より，強化学習はマルコフ性を仮定するために行ったり来たりといった余計な挙動を挟むが，上記の

設定においては初ゴールまでは大きく経路を変更するのに対し，それ以降は経路変更が小さくなり少ない回数で学習が完了することも分かった．第 4.5.3 項の結果と照らし合わせると，この挙動は人の経路選択の挙動と一致する部分がある．これらの点から考えると，3 手更新の Multi-step Learning で学習率を 0.9 程度まで高めた設定は，人に比較的近い学習挙動をしていると捉えることができるだろう．しかし，学習率が高いと最適な解に収束しない可能性が高まるという一般的な性質がある．今回の実験環境においてはゴールまでの解が 1 つしかないため最短経路に学習を収束させることができているが，例えば，最短経路の他に複数の最短でない解が存在するような迷路においては上手く学習が進まないことが想定される．そこで，そのような迷路における人の学習過程を追加で調査する必要があるだろう．

7. まとめ

本研究では人の迷路学習過程を調査するために，VR 空間内に強化学習の環境を模した迷路を作成した．その環境を用いて人に対して迷路探索課題を行い，人の迷路学習過程を動物迷路実験の知見や強化学習の枠組みを絡めて考察をした．結果として，人は強化学習のような手順を覚える学習法と認知地図のような空間情報を活用した学習を組み合わせている特徴が見られた．これにより，強化学習の学習速度と比べ非常に速く，かつ，形状が微小量変化した迷路に関しても一度学習した結果を活かして効果的に学習を進めていると結論付けた．また，既存の強化学習手法について，更新間隔が 3 手の Multi-step Learning の学習率を 0.9 程度まで高めた設定にすると，人の行動に比較的近い挙動を再現することができた．しかし，本研究システムにはまだ改善の余地があり，かつ検証する余地のある設定が残っていることが課題としてある．

参考文献

[1] 田熊一敞, 永井拓, 山田清文. (2007). 学習・記憶行動の評価法. 日本薬理学雑誌, 130(2), 112-116.
[2] Morris, R. G. M. (1981). Spatial localization does not require the presence of local cues. Learning and Motivation, 12(2), 239-260. doi: https://doi.org/10.1016/0023-9690(81)90020-5
[3] Skinner, B. F. (1963). Operant behavior. American Psychologist, 18(8), 503.
[4] 吉本潤一郎, 伊藤真, 銅谷賢治. (2013). 《第 10 回》 脳の意思決定機構と強化学習. 計測と制御, 52(8), 749-754.
[5] Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. Science, 275(5306), 1593-1599.
[6] Shidara, M., Aigner, T. G., & Richmond, B. J. (1998). Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. Journal of Neuroscience, 18(7), 2613-2625.
[7] Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. Science, 310(5752), 1337-1340.
[8] A.G. Barto, Adaptive critics and the basal ganglia, in Models of Information Processing in the Basal Ganglia, J.C. Houk, J. Davis, and D. Beiser(eds.), pp.215-232, MIT Press, 1995.
[9] Doya, K. (2002). Metalearning and neuromodulation. Neural Networks, 15(4-6), 495-506.
[10] Rasmusson, D. D. (2000). The role of acetylcholine in cortical synaptic plasticity. Behavioural Brain Research, 115(2), 205-218.
[11] Buhot, M. (1997). Serotonin receptors in cognitive behaviors. Current Opinion in Neurobiology, 7(2), 243-254.
[12] Rahman, S., Sahakian, B. J., Cardinal, R. N., Rogers, R. D., & Robbins, T. W. (2001). Decision making and neuropsychiatry. Trends in Cognitive Sciences, 5(6), 271-277.
[13] 齋藤智輝, 大枝真一. (2011). 動的環境を対象とした適応的強化学習の提案. 第 73 回全国大会講演論文集, 2011(1), 205-206.
[14] 寺井輝, 小宮山摂.(2020). 動的環境を有する迷路を対象とした強化学習における学習モデルの再利用. 2020 年電子情報通信学会学生ポスターセッション, ISS-SP-052, 2020(5)
[15] Tolman, E. C. (1948). Cognitive maps in rats and men. Psychological Review, 55(4), 189.
[16] 北濱亨, 三浦利章, 岡崎甚幸, 篠原一光, 田村仁志, 松井裕子. (1999). 迷路探索歩行時の注視と歩行に関する研究. 人間工学, 35(3), 145-155.
[17] 佐々木隆宏, 阪口豊, 出澤正徳, 小宮山摂. (2005). 強化学習型タスクにおける人間の行動決定に関する研究. 電気通信大学大学院 情報システム学研究科修士論文.
[18] 小川一仁. (2009). 強化学習モデルは人間行動をどの程度説明するか?: 均衡が 1 度だけ移動する経済実験を例に. 經濟論叢, 183(3), 59-71.
[19] WATKINS, C. J. C. H. (1989). Learning form delayed rewards. Ph.D.Thesis, King's College, University of Cambridge, Retrieved from https://ci.nii.ac.jp/naid/10007782517/
[20] Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. Machine Learning, 3(1), 9-44.
[21] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction MIT press. 141-158.