代理強化における観察対象の熟練度・類似度がエージェントの学習にもたらす影響の評価

Evaluating the Effects of Proficiency and Similarity of Observed Objects on Agent Learning in Vicarious Reinforcement

原田 雄大 *　　竹内 勇剛
Yudai Harada　　Yugo Takeuchi

静岡大学
Shizuoka University

Abstract: 人間は多様な行為主体が存在する環境においても他者の振舞いを見て学習することで，素早く適応できている．実社会において行為主体は身体・物理的な特徴があり，求められる振舞いが異なってくるにも関わらず，属性の異なる他者から模倣学習を行うことができている．しかし，このように人間が他者を模倣し学習する認知メカニズムは明らかになっていない．そこで本研究では人間が行う模倣の構造を明らかにするための足掛かりとして，本稿では，模倣対象である行為主体の属性と熟練度が学習の効率化に寄与するのか，マルチエージェントの学習シミュレーションによって検証した．実験の結果，模倣対象の最適化された行動方策より，学習途中の行動方策を模倣したほうが無駄なく学習できることが示唆された．また，属性が異なっていても模倣対象が最適行動を内包している場合，効率的な模倣が行えることが示唆された．

1　はじめに

人工知能や機械学習といった機械の知能化・自動化が進む中，現実社会における機械のあり方は変化してきている．最近では人間が行ってきたコミュニケーションを伴う業務や支援を機械に代替させる動きが増えている [1]．このように，これからの機械は人間から道具として扱われるだけではなく，個人や集団といった動的に変化する複雑系の中で環境を共有し，目的を遂行する必要がある．状況が目まぐるしく変化する人間社会において，人間は複雑な環境や未知の課題に対したとき，素早く適応することが可能である．これまで機械が人間のように複雑な環境の中でも素早く適応できるように，様々な角度から行動方策の構築，モデル化が行われてきたが，適応できる状況が限られ汎用性に欠けることや複雑な問題を最適化するまでのコストが膨大になることなど課題があり，人間のように未知の環境においても素早く学習し，複雑な環境に適応する汎用的な機械の実現には至っていない．そこで，本研究では人間のように素早く適応し振る舞う機械を実現するためにはどうしたらよいかという問題を研究課題とした．

これまで，人間の振舞いをコンピューターシミュレーションによって再現し，理解するといった研究は数多く行われてきた [2][3]．人間と機械やロボット，エージェントのインタラクションを研究する Human-Agent Interaction の分野では，決められた状況の中での応答に対してエージェントに行動モデルを付与し振る舞わせることで人間が持つ社会性を見出せることを示唆している [4][5]．これらの研究はエージェントに対し，トップダウンに行動モデルを与えることで人間が持つ社会性をエージェントに付与している．このようなトップダウンにモデルを構築する手法は，エージェントの行動方策や環境内のルールを事前に知識として与える必要がある．これらの行動方策やルールは相対する環境や課題に依存するため，他の環境や課題に応用することが難しい．多様なシチュエーションが起こりえる人間社会においてトップダウンのモデル化は汎用性に欠ける [6]．一方で，ボトムアップなモデル構築手法はモデルの枠組みだけを設計するだけで機械がモデルの内部を構築してくれるため多様な状況をあらかじめ想定することなくモデルの構築を行うことができる [7]．

ボトムアップなモデル構築手法の一つに強化学習がある．強化学習は行動主体であるエージェントと環境とのインタラクションを経験として蓄積し，経験を基に個体レベルの振る舞いをボトムアップにモデル化する手法である．設計者は細かいインタラクションのシチュエーションを想定する必要がなくエージェントは

*連絡先：静岡大学大学院総合科学技術研究科
〒 432-8011 静岡県浜松市中区城北 3-5-1
E-mail: harada.yudai.19@shizuoka.ac.jp

環境や与えられたルールの範囲内でボトムアップに行動方策を獲得できるため，人間社会での多様なシチュエーションに対応できるモデルを構築可能である [8]. Nagata et al.(2007) や保田ら (2013) は強化学習を用いた研究を行っており，人間社会を想定した協調課題に対してマルチエージェントの強化学習を行い，集団内でエージェントが社会的なインタラクションが創発することを示している [9][10]．また，現実社会のインタラクションを考えたとき，人間が日常的に行っている現実空間での身体的なインタラクションであったり，仮想空間における身体性を排したテキストベースのインタラクションであったり，人間は様々なインタラクションを行っている．本研究では身体をインターフェイスとして扱うエージェントが身体的インタラクションの過程で人間のように素早く適応し振る舞うことに着目するため，身体的・物理的レベルでのインタラクションを行う行動モデルの構築が必要である．近年の強化学習はマルチエージェントのゲーム課題や物理演算シミュレーションに適したアルゴリズムの開発も行われており，身体性を有した現実空間に近い環境でのシミュレーションにも適している [11]．しかし，これまで強化学習を用いた研究の多くは単一の属性で構成された行為主体が複数存在する簡単な課題におけるインタラクションについて言及しており，人間社会のように多種多様な行為主体が存在する複雑環境でのインタラクションについては十分に研究されていない．多種多様な行為主体が存在する環境において個体間の想定されるインタラクションは膨大となり，計算量は増加する．また，膨大な状態数が想定される環境で最適化するとなると必要となるモデルも複雑になり，設計が難しくなるといった問題もでてくる [12].

人間が環境に適応するまでの学習に目を向けると，一度学習した内容を他の状況にも活かすといった転移学習や他者の行動や振舞いを見て学習するといった模倣学習があり，このような学習を日常的に行うことで多種多様な行為主体が存在する複雑な環境下でも素早く適応していると考えられる [13]．そこで本研究では人間が環境や課題に適応するまでのプロセスである学習に着目する．人間は日常的に模倣や転移を繰り返し行っていると考えられるが，どのような条件・過程でこれらが起こり最適化を行っているのか明らかになっていない．このような学習のメカニズムを解明するためには，人間の意思決定と同様にボトムアップに行動方策を確立する環境で構成論的に理解する必要がある．そこで人間が行う模倣の構造を明らかにするための足掛かりとして，本稿では，模倣対象である行為主体の属性と熟練度が学習の効率化にどのように影響するのか，マルチエージェントの強化学習シミュレーションによって検証した．

具体的には，多様な行為主体が存在する環境での移動課題を題材にして，マルチエージェント強化学習を行い，各エージェントの行動の最適化を行う．模倣を行う学習者には事前に学習した他者の行動方策を与え，模倣学習を行わせる．この学習実験を通して，多種多様な行為主体が存在する環境で最適化するための効率的な学習モデルについて議論する．多種多様なエージェントが混在する複雑環境での効率的な学習手法が明らかになれば，複雑な環境に適応する機械やエージェントを実現するための学習手法を示すことができる．本研究の成果は，人間社会のように多種多様な行為主体が存在する環境における，人間やエージェントなどの行為主体の学習モデルのデザインに寄与する．また，複雑空間での課題に対した最適化における効率的な学習手法の確立も期待できる．

2　多様な行為主体が混在する移動課題における学習

2.1　人間の学習

人間の学習には他者の振舞いを模倣し，学習する模倣学習がある [13]．模倣学習といっても様々あり，Flanders は学習者が直接環境内で模倣対象とインタラクションし学習する「直接強化」と学習者は模倣対象の学習を環境外や別環境から観測し，学習する「代理強化」と混在した環境や連続した環境で双方の手法によって学習する「2 重強化」の三つに分類されると述べている [14]．また，春木は模倣対象を手掛かりにして対象と同一反応することの学習である「模倣の学習」と模倣対象と同一反応することによって環境の手掛かりを習得する「模倣による学習」の二つに分類されると述べている [15]．前者の「模倣の学習」は模倣することの学習と同義であり，学習することの学習を行う「メタ学習」[16] に近いものである．また,「模倣による学習」は他者の振舞いや経験を基に自身が学習することであるので,「観察学習」[13]「転移学習」[17] と同等の学習である．このように人間は効率よく最適化するために，学習する術を持っている．

2.2　機械の学習

機械に人間の学習手法を導入した一つとして強化学習に模倣を取り入れたアルゴリズムがある．この模倣学習は最適化した行動や人間の振舞いをデモとして与え，機械に成功体験を事前に学習させる [18]．その後，自身の最適化を強化学習に切り替えて行うといった手法である．このアルゴリズムは最適化した行動方策やその課題に対するエキスパートの行動方策を事前に与えることで学習初期の探索を省き，効率よく学習させる手法であるが，最適化した行動方策が事前にない課題や模倣対象と学習者の属性が大きく異なる場合適応

できないことがある．一方で人間社会において，主体となる人間は特徴・属性は千差万別であり，模倣対象となる他者は特徴や属性が異なることが多いが人間は模倣を行い学習することが可能である．このような模倣学習を機械に導入するには属性の異なる他者のどのような知識や行動を模倣し学習しているのか明らかにする必要がある．

そのため，本研究では「模倣による学習」の過程で属性の異なる他者からどのように模倣しているのか明らかにするためにエージェントシミュレーションを用い，模倣学習を行う．模倣学習を行わせる際，模倣対象とともにインタラクションを行う「直接強化」では，効率的な模倣が創発する要因を検討することが難しい．したがって，他者の経験や知識を転用する「代理強化」による模倣を用いて，どの要因が属性の異なる他者からの効率的な模倣に寄与するのか確かめる必要がある．

2.3 移動課題のルール

本研究では多様な行為主体が存在する環境での課題として複数の行為主体での移動課題を扱う．本稿における移動課題とは図 1 のように環境内にエージェントと障害物を用意し，エージェントは障害物を避けながらランダムに指定された場所まで移動する課題である．エージェントは他エージェントと接触したり，指定時間内に目的地まで移動できなかった場合は課題失敗となる．本課題は複数のエージェントと同時に課題を行うマルチエージェントタスクであり，すべてのエージェントが同時に課題をはじめ，すべてのエージェントが終了条件（目的地まで移動したか課題に失敗したか）になるまでを 1 試行とする．エージェントは図 1 のように視線を有しており，自身の近くにあるオブジェクト（障害物，他エージェント）の位置を把握することが可能になっている．

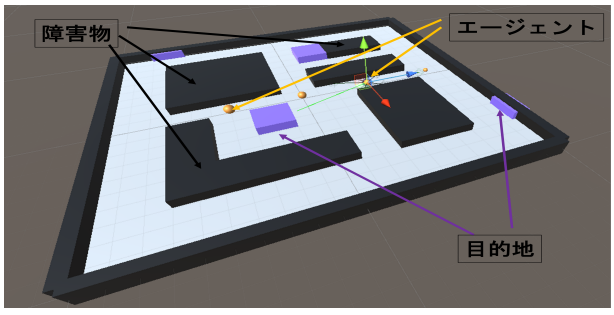


図 1: 移動課題の環境.

このような多様な行為主体が混在する移動課題は，複雑環境での身体的なインタラクションについて観察・評価できる．また，与える身体に実際に考えられる身体的特徴を付与すれば行為主体を多様にすることも容易である．このように多様な行為主体を設計して学習すれば，各行為主体によって最適な行動方策は変化してくると考えられる．人間は行動方策の異なる他者からも模倣し学習することができると考えられるが，そのメカニズムは明らかになっていない．そこで，本研究では代理強化として事前に与える知識によって，行われる学習に変化があるのか，移動課題を用いて検証する．

3 強化学習手法

3.1 強化学習

本研究では，ボトムアップにモデルを構築する手法に強化学習を用いる．強化学習とは行動主体が環境の中で与えられる報酬と状態をもとに行動を繰り返し，方策を更新していき最適化を行うものであり，行動主体であるエージェント自身が経験し行動のモデルを構築するためボトムアップにモデルを構築することができる．代表的な学習法としては Q 学習 [19] や Q 学習にニューラルネットワークを用いた DQN[20] やゲーム課題に用いられることがある PPO[11] などが存在する．本研究での移動課題は実社会への応用を考えているため 3 次元環境で行う．Q 学習の状態空間は離散値を扱っているため適していない．DQN の状態空間は連続値を扱うが出力である行動空間は離散値しか表現できない．一方で PPO では入力である状態空間と出力である行動空間の両方を連続値であるかうことができるため 3 次元環境での学習に適している．さらに PPO は DQN に比べ環境の変化に対する頑健性を持っているためマルチエージェントでの学習に適していると考えられる．以上の観点で強化学習手法を比較した結果を表 1 に示す．表 1 より 3 次元空間におけるマルチエージェントでの移動課題は PPO が適していると結論付けた．

表 1: 強化学習手法の比較

手法	連続空間	状態数	環境の変化
Q 学習	×	×	×
DQN	△	○	×
PPO	○	○	○

3.2 PPO

Unity ML-agents に搭載されている PPO（Proximal Policy Optimization）は環境からの情報取得と目的関数の最適化を交互に繰り替えるアルゴリズムであり，マルチエージェントのゲーム課題や物理演算シミュレーションに適したアルゴリズムである [11][21]．PPO に

おける方策の更新は式 1 を目的関数とした勾配法を用いる．PPO の特徴は方策の更新を行うときに変化量が大きくならないようクリッピングを行う点である．クリッピングは式 1 中の clip 関数で行っており，式 2 に示す方策の変化の比が指定した値を超えた場合に変化量を一定の値にする．このクリッピングによって前の方策からの偏差を抑え，計算の複雑さを減らすことで学習の安定化を行っている．式 1 の θ は方策パラメータ，$\hat{\mathbb{E}}_t$ は期待値，$\hat{A}_t$ は Advantage の推定値，ϵ はクリッピングの大きさを示している．図 2 には PPO のアルゴリズムのフローチャートを示す．

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{1}$$

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{2}$$

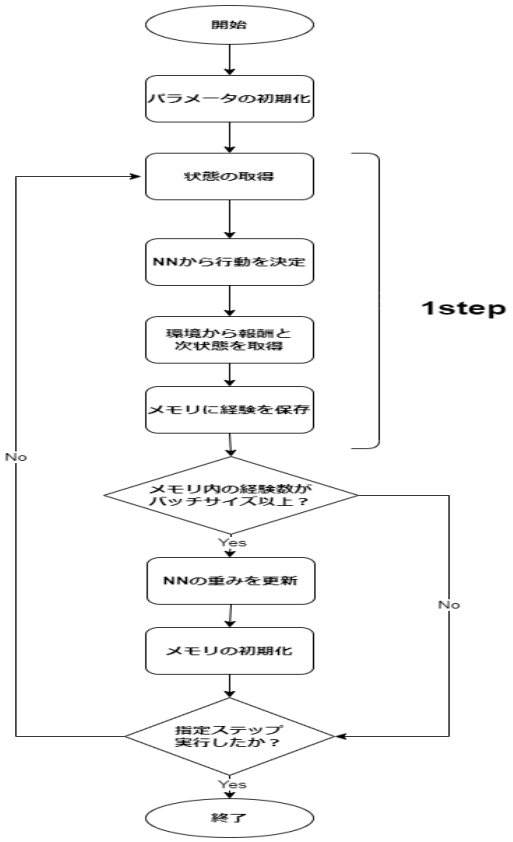


図 2: PPO のフローチャート

3.3 各種パラメータ

強化学習ははじめに状態空間，行動空間，報酬と学習率や割引率などのハイパーパラメータを決定して学習を始める．本節では移動課題でのエージェントの状態空間，行動空間，報酬系についての説明と PPO のハイパーパラメータの値を示す．移動課題の強化学習におけるエージェントの状態空間はエージェントの位置情報と速度，目的地の座標に加え，視線情報から構成される．内容や次元数をまとめたものを表 2 に示す．

表 2: 状態空間

状態	次元数
エージェントの絶対位置座標	2
エージェントの速度	2
ゴール地点の絶対位置座標	2
視覚情報（オブジェクトとの距離)）	27

本課題におけるエージェントの行動は目的地に向かうための移動である．今回は 3 次元仮想環境の中で縦方向と横方向に力を加えることで移動するように設計した．したがって，縦方向と横方向の 2 次元の連続値を出力するように設定した．また，報酬系は表 3 のように細かいイベントに対して報酬とペナルティを細かく与えるように設計した．

表 3: 報酬設計

イベント	reward・penalty	値
目的地到達	r	1.0
時間経過	p(毎ステップ)	0.0001
障害物接触	r・p(毎ステップ)	0.0001
停止時間	r・p(毎ステップ)	0.0001
急加減速	r・p(毎ステップ)	0.0001
移動距離	r・p(毎ステップ)	0.0001
他者接触	p	0.7
時間切れ	p	0.7

PPO におけるハイパーパラメータは表 4 のように設定した．

表 4: PPO のハイパーパラメータ

パラメータ名	値
バッチサイズ	2024
バッファサイズ	51200
方策変化量の閾値 ϵ	0.2
エントロピー正規化率 β	0.001
正規化パラメータ λ	0.95
学習率 η	0.0003
割引率 γ	0.995
エポック数	3
隠れ層のニューロン数	256
隠れ層の数	3

4　学習実験

4.1　実験環境

本研究では，図 1 のような仮想環境を Unity 上に構築し，エージェントが行動方策をボトムアップに構築する手法である強化学習を用いた学習実験を行う．学習は Unity の強化学習ライブラリである ML-Agents を用いて行う．本実験で使用したハードウェアおよびソフトウェア環境を表 5 に示す．

表 5: 開発環境

種別	名称 (備考)
OS	Windows 11(64bit)
プロセッサ	Intel(R) Core(TM) i7-8700 CPU
RAM 量	32[GB]
ゲーム開発ライブラリ	Unity (Ver.2020.3.30f1)
強化学習ライブラリ	ML-Agents (release18)

4.2　実験条件・評価方法

多種多様な行為主体が存在する環境で模倣学習を行うとき，効率的な学習に適している条件を明らかにするために，本実験では「属性の類似段階」と「模倣対象の学習段階」の 2 要因を用いる．「属性の類似段階」とは模倣対象である行為主体と学習者の属性の類似度である．ここでの属性とは行為主体に与える身体的特徴である．エージェントには身体的特徴として大きさ，最大速度，加速度の 3 変数を変化させ与えているため，各エージェントごとに属性が変化してくる．模倣者と模倣対象の属性がより離れていると，一方の最適行動ともう一方の最適行動は異なってくる．そのため，模倣対象の知識の転用が難しく学習を効率的に行うことができないと考えられる．そのため，本実験では学習者の属性を段階的に変化させ，模倣対象となるエージェントとの類似度を変え，それぞれ模倣学習を行わせる．属性の変化にはエージェントの大きさを用い，図 3 のように模倣対象となるエージェントの大きさを基準にして，小さくしたものを 2 条件 (XS，S)，大きくしたものを 2 条件 (XL，L) 用意した．

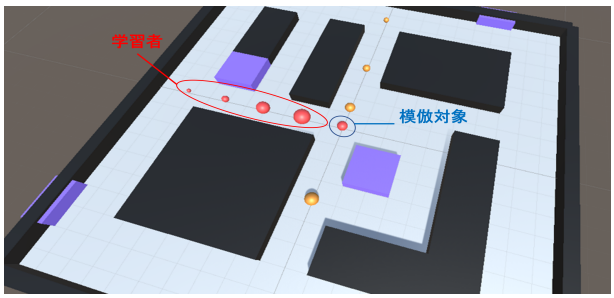


図 3: エージェントの属性 (大きさ) の変化.

表 6: 実験条件

条件名	属性の類似段階	模倣対象の学習段階
XS-2	-3.0(XS)	2Mstep 時 (学習初期)
XS-4	-3.0(XS)	4Mstep 時 (学習中盤)
XS-8	-3.0(XS)	8Mstep 時 (学習終盤)
XS-50	-3.0(XS)	50Mstep 時 (最適化済)
S-2	-1.5(S)	2Mstep 時 (学習初期)
S-4	-1.5(S)	4Mstep 時 (学習中盤)
S-8	-1.5(S)	8Mstep 時 (学習終盤)
S-50	-1.5(S)	50Mstep 時 (最適化済)
L-2	+1.5(L)	2Mstep 時 (学習初期)
L-4	+1.5(L)	4Mstep 時 (学習中盤)
L-8	+1.5(L)	8Mstep 時 (学習終盤)
L-50	+1.5(L)	50Mstep 時 (最適化済)
XL-2	+3.0(XL)	2Mstep 時 (学習初期)
XL-4	+3.0(XL)	4Mstep 時 (学習中盤)
XL-8	+3.0(XL)	8Mstep 時 (学習終盤)
XL-50	+3.0(XL)	50Mstep 時 (最適化済)

「模倣対象の学習段階」とは模倣対象の最適化がどのレベルで行われているかである．模倣対象の最適化が完全に終わっていると，その行動方策を利用しても，属性の異なる自身で利用しても最適な振舞いではなく，最適化が行われない，もしくは再度学習する必要があると考えられる．一方で学習がある程度進み，環境に慣れた状態の観察対象の行動方策を利用すると，属性が異なっていても，環境に対する初期の学習 (慣れ) だけを模倣し，その後の学習で自身の最適な振舞いを効率よく学習することができると考えられる．本実験では他者の知識を転用し学習を行うため，模倣対象であるエージェントはあらかじめ知識を保持しておく必要がある．そのため，模倣対象となるエージェントは同環境であらかじめ学習し，最適化を行ったエージェントとしする．その学習過程の行動方策を知識として与えることで模倣学習を行わせる．事前に行った模倣対象となるエージェントの学習結果を図 4 に示す．

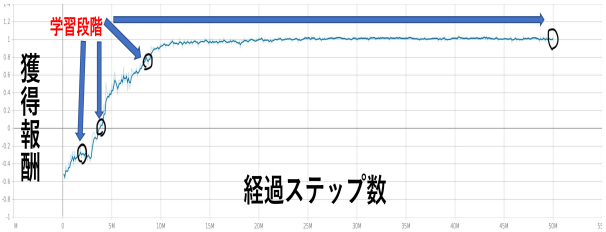


図 4: 模倣対象の事前学習.

図 4 のように学習初期，中盤，終盤，学習収束と過程があるため，どのタイミングの知識を転用するかを条件として扱う．今回の実験では 2Mstep 時 (学習初期),

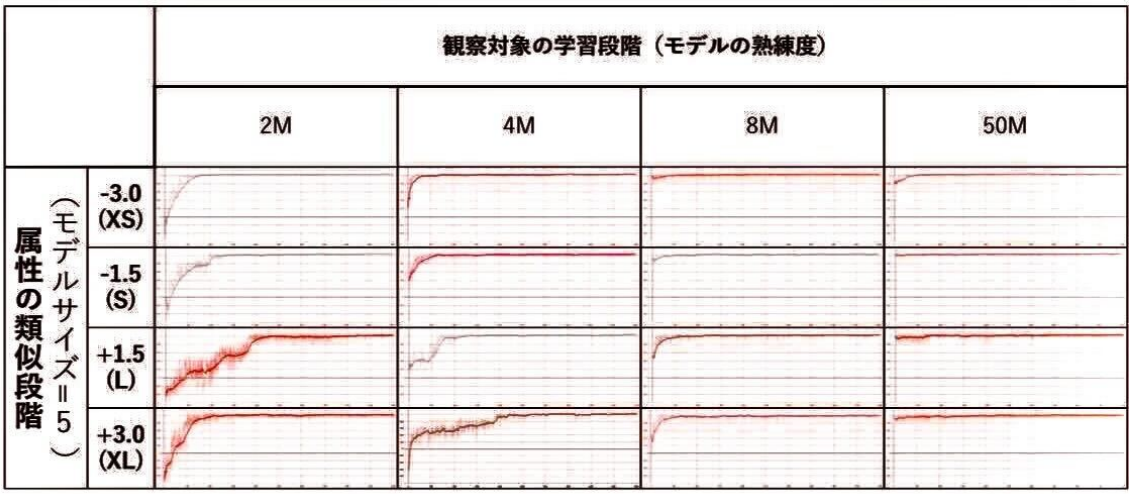


図 5: 16 条件の獲得報酬遷移.

4Mstep 時 (学習中盤), 8Mstep 時 (学習終盤), 50Mstep 時 (学習収束) の行動方策 (ニューラルネットワーク) を学習者となるエージェントの初期値に与えることで代理強化を行う．このように模倣対象である他者の属性と学習段階は模倣学習に影響を及ぼすと考えられる．また，属性の違いによって最適な模倣を行える学習段階は変化してくると考えられる．そこで本研究では表 6 のように「属性の類似段階」と「模倣対象の学習段階」の 2 要因を用いた 16 条件で学習実験を行う．

各条件においてエージェントは 5 体用意する．5 体のうち 4 体はすでに学習し，最適化したエージェントを用い，残りの 1 体は事前学習して最適化を行った模倣対象であるエージェントの知識 (行動方策) を有したエージェントである．この 5 体のエージェントで移動課題を同環境で行う，マルチエージェント強化学習を 50Mstep 行った．各条件における模倣対象が学習する過程で最適化までにかかるコスト (学習時間) の観点から学習効率を評価し，模倣学習を効率的に行うために必要な条件を明らかにする．

4.3 実験結果

PPO を用いて各条件で 50M ステップ学習した結果を図 5 に示す．なお，図 5 には模倣学習を行ったエージェント 1 体の獲得報酬遷移を示しており，横軸はエージェントの学習ステップ数で縦軸はエージェントの平均獲得報酬を示している．

まず，模倣対象の学習段階に着目し，実験結果を整理する．図 6a に属性が S の 4 条件，図 6b に属性が L の 4 条件，図 6c に属性が S の 4 条件，図 6d に属性が S の 4 条件で学習したときの獲得報酬報酬遷移をまとめたものを示す．図 6a，6b を見ると，模倣対象と属性の近い S の 4 条件・L の 4 条件は学習が進んだ模倣対象の行動方策を利用することで最適化が早く行われている．図 6c を見ると身体的大きさが一番小さい XS の 4 条件では，学習が進んだ模倣対象の行動方策を利用しても，2M 条件以外は学習が収束するまでのステップ数は変化がなかった．特に 8M，50M 条件での学習過程に着目すると，学習初期に獲得している報酬が逆転している．50M 条件では一度最適化を行っているため，獲得できる報酬は多いはずだが，8M 条件の獲得報酬のほうが高くなっている．図 6c から身体的大きさが一番大きい XL の 4 条件を比較すると，模倣対象の最適化が中盤の 4M 条件の学習進度が遅いことがわかる．一方で，8M，50M 条件では素早く学習が進んでおり，効率的な学習が行えたといえる．

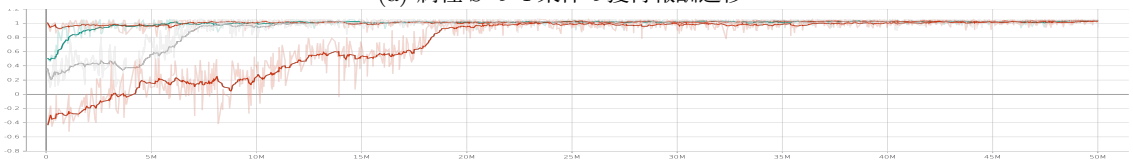
次に，模倣対象の類似段階に着目し，実験結果を整理する．類似度に関しては，学習者と模倣対象の属性が離れているとより学習進度が遅くなると予測していたが，図 5 を見ると S-2 条件,S-4 条件に比べ，XS-2 条件，XS-4 条件のほうに早く学習できている．また，L-2 条件や XL-4 条件は学習が進むのが遅く最適化までに時間を要している．

4.4 考察

4.4.1 熟練度がもたらす学習への影響

エージェントの大きさが一番小さい XS の 4 条件では，学習初期の行動方策を模倣する 2M 条件以外は学習が収束するまでのステップ数は変化がなく，特に 8M，50M 条件での学習過程に着目すると，学習初期に獲得している報酬が逆転した．50M 条件では一度最適化を

(a) 属性 S の 4 条件の獲得報酬遷移



(b) 属性 L の 4 条件の獲得報酬遷移

(c) 属性 XS の 4 条件の獲得報酬遷移

(d) 属性 XL の 4 条件の獲得報酬遷移

図 6: 学習の結果 16 条件でエージェントが獲得した報酬の推移

行っているため，獲得できる報酬は多いはずだが，模倣対象の最適化が終盤の 8M 条件の獲得報酬のほうが高くなっている．これは最適行動の異なる他者から模倣を行っているからだと考えられる．このことから属性が大きく異なると最適行動も変化するため，一度最適化を終えた知識を転用すると，かえって非効率的になることが示唆された．エージェントの大きさが一番大きい XL の 4 条件においては，4M 条件の学習進度が遅い一方で，8M，50M 条件では素早く学習が進んでいる．今回は獲得報酬遷移からエージェントの最適行動の獲得と学習効率について言及しているため，学習過程のエージェントの振舞いは評価できていないが，最適化中盤を模倣する 4M 時の行動方策が XL の属性をもつエージェントには不利益なものであると考えられる．これらの結果は，学習段階（行動の熟練度）が単純に影響しているのではなく，学習途中で創発される行動方策が異なる属性をもつエージェントにとっては学習を阻害する可能性があることを示唆している．

4.4.2 属性の変化がもたらす学習への影響

類似度に関しては，属性として与えたエージェントのサイズが小さくなるに連れて早く学習できているが，大きくなると学習が進むのが遅く最適化までに時間を要している．この結果は行動方策が内包しているか否かが関係していると考えられる．今回の実験ではエージェントの大きさを変化させ属性を与えているため，本移動課題においてより大きなエージェントは接触の危険が大きくなることが想定される．つまり，小さなエージェントが通れるルートを大きなエージェントは通ることができない場面が存在しうる．しかし，大きなエージェントが通れるルートは小さなエージェントも通れることが多い．このように属性によっては適した行動が内包されていることがあるため，属性が離れていても模倣し効率よく学習することが可能であることが示唆された．

以上の考察をまとめたものを図 7 に示す．図 7 には観察対象の属性が学習者と類似しているかによって模倣しやすくなるか，また模倣対象の熟練度によって模倣しやすくなるかを示している．

これらの知見は，機械が最適行動を獲得するまでの効率的な学習モデルに寄与する．しかし，本研究では 1 次元の属性でしか類似度を変化させていない．属性が異なっていても最適行動が内包される場合模倣を行い効率的な学習を行えることが示唆されていることから，ある変数において属性が異なっていても，別の次元では類似している場合模倣し学習することができると考えられる．以上のことから，異なる属性から類似している箇所を見つけ出し，最適行動の抽出し模倣を行うことができるエージェントの確立を期待できる．

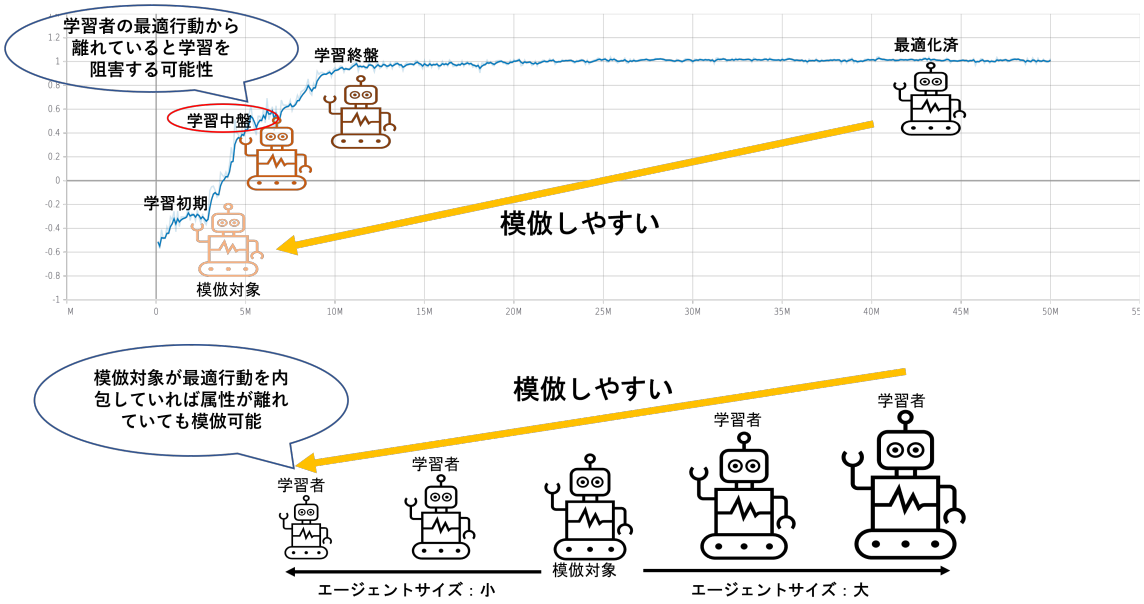


図 7: 本研究で明らかになった模倣の構造.

5　まとめと今後の計画

本研究では人間が行う模倣学習に着目し，エージェントに同様の学習を実現するために必要な条件や認知過程について構成論的に検証した．具体的には身体をもつ多様なエージェントを用いて，ボトムアップなモデル構築手法である強化学習シミュレーションを行った．エージェントに与える模倣対象の属性や熟練度を変化させ，学習過程や創発される振舞いを分析することで，どの要因がエージェントの効率的な学習に寄与しているのか検証した．実験結果として以下の二つが示唆された．

(1) 模倣対象である他者の最適化が行われ，属性が大きく異なる場合，他者の行動方策を転用できず非効率になることが示唆された．しかし，学習中盤の行動方策を与えて学習を行ったとき，最適化をスムーズに行えなかったり，学習が滞ったりしたことから，模倣対象の学習過程で創発した行動方策が学習を阻害する可能性があることを示唆している．

(2) 属性が大きく異なるエージェントに行動方策を転用しても効率よく学習することができており，属性が離れていても最適行動が内包されていることがあるため，模倣し効率よく学習することが可能であることが示唆された．

これらの結果はエージェントが効率的な学習を行う条件について言及しており，エージェントの新しい学習モデルの構築に寄与する．また，実環境に近い連続空間においてエージェントが模倣 (代理強化) を行い，身体的インタラクションの過程で最適化する学習シミュレーションを用いて検証実験を行ったため，複雑環境の中で人間が行う模倣行動の認知メカニズムの解明に寄与し得る．本研究の限界として，事前に学習したエージェントの行動方策をトップダウンに与え，模倣を再現しており，本実験ではトップダウンの模倣学習 (代理強化) についてしか言及できていない．模倣はいくつかに分類されており，直接強化や模倣の学習といった模倣については議論できていない．学習者が環境内で模倣対象と直接インタラクションを行い，模倣する直接強化はインタラクションの過程で模倣対象や模倣箇所を見つける必要があるためボトムアップに行動し学習していると考えられる．このようなボトムアップな模倣を機械に行わせるためにはどのように模倣行為を認知し，どのように学習するのかといったメタ学習のメカニズムを明らかにする必要があるため，今後は，模倣行為の学習という観点からエージェントモデルの検証を行い，ボトムアップな模倣のメカニズムの解明を行っていきたい．

参考文献

[1] 大澤博隆, Joseph F.Coughlin, 今井倫太, 山田誠二: 擬人化表現を介した米国の高齢者に対する情報提示手法の開発, 情報処理学会研究報告, Vol.2010, No.4, pp.1-8(2010).

[2] Smith, R.G., & Davis, R.: Frameworks for cooperation in distributed problem solving, IEEE Transactions on systems, man, and cybernetics, Vol.11, No.1, pp.61-70(1981).

[3] Durfee, E.H., Lesser, V.R., & Corkill,D.D.: Trends in Cooperative Distributed Problem Solving, IEEE Transactions on Knowledge and Data Engineering, Vol.1, No.1, pp.66-71(1989).

[4] 竹内勇剛, 片桐恭弘: ユーザの社会性に基づくエージェントに対する同調反応の誘発, 情報処理学会論文誌, Vol.41, No.5, pp.1257-1266(2000).

[5] 中西英之, 石田亨: 仮想空間内でのコミュニケーションを補助する社会的エージェントの設計, 情報処理学会論文誌, Vol.42, No.6, pp.1368-1376(2001).

[6] Sen,S. & Sekaran, M.: Multiagent coordination with learning classi

er systems, In International Joint Conference on Arti cial Intelligence, pp.218-233(1995).

[7] 山村雅幸, 宮崎和光, 小林重信: エージェントの学習, 人工知能学会誌, Vol.10, No.5, pp.683-689(1995).

[8] 内田英明, 藤井秀樹, 吉村忍, 荒井幸代: 道路ネットワークの変化に対する経路選択の学習, 情報処理学会論文誌, Vol.53, No.11, pp.2409-2418(2012).

[9] Nagata, Y., Ishikawa, S., Omori, T., & Morikawa, K.: Computational model of cooperative behavior: Adaptive regulation of goals and behavior, Proceedings of the European Cognitive Science Conference, pp.202-207(2007).

[10] 保田俊行, 大倉和博: 連続空間における強化学習によるマルチロボットシステムの協調行動獲得, 計測と制御, Vol.52, No.7, pp.648-655(2013).

[11] JunLAI, Xi-liang, C., & Xue-zhen, Z.: Training an Agent for Third-person Shooter Game Using Unity ML-Agents, Intemational Conference on Artficial Intelligence and Computing Science (ICAICS 2019), pp.305-310(2019).

[12] 荒井幸代: マルチエージェント強化学習: 実用化に向けての課題・理論・諸技術との融合, 人工知能学会誌, Vol.16, No.4, pp.476-481(2001).

[13] Bandura, A. & Aletha, C. H.: Identification as a process of incidental learning, The Journal of Abnormal and Social Psychology, Vol.63, No.2, pp.311-318(1961).

[14] Flanders ,J.P.: A review of research on imitative behavior, Psychological Bulletin, Vol.69, No.5, pp.316–337(1968).

[15] 春木豊, 都築忠義: 模倣学習に関する研究, 心理学研究, Vol.41, No.2, pp.90-106(1970).

[16] 横山拓, 鈴木宏昭: 洞察問題解決におけるメタ学習, 認知科学, Vol.25, No.2, pp.156-171(2018).

[17] 白水始: 認知科学と学習科学における知識の転移, 人工知能, Vol.27, No.4, pp.347-358(2012).

[18] Ciosek, K.: Imitation learning by reinforcement learning, arXiv preprint arXiv, 2108.04763.(2021).

[19] Amy, G. & Keith, H.: Correlated-Q Learning, ICML, pp.84-89(2003).

[20] Yingfeng, C., Shaoqing, Y., Hai, W., Chenglong, T., & Long, C.: A Decision Control Method for Autonomous Driving Based on Multi-Task Reinforcement Learning, IEEE Access2021, pp.154553-154562(2021).

[21] Bøhn, E., Coates, E. M., Moe, S., & Johansen, T. A.: Deep reinforcement learning attitude control of fixed-wing uavs using proximal policy optimization, International Conference on Unmanned Aircraft Systems (ICUAS 2019), IEEE, pp.523-533(2019).