集団は多様な価値観のもとで秩序化されているのか
Are Groups Ordered by Diverse Values

原田 雄大† 竹内 勇剛†
Yudai Harada & Yugo Takeuchi
† 静岡大学 大学院総合科学技術研究科
Graduate School of Integrated Science and Technology, Shizuoka University
harada.yudai.19@shizuoka.ac.jp

概要

人間は多種多様な他者が存在するような複雑環境において，状況に応じて価値観を変化させ円滑なインタラクションを図ることがある．一方で複雑な環境で円滑なインタラクションを行える機械は限定的な状況にしか適応できず，多種多様な行為主体が存在する環境に適応できるモデルは少ない．本研究では，円滑にインタラクションを行える汎用的なエージェントモデルを実現するために，人間のように価値観を動的に変化させる振舞いを強化学習によって獲得できるか確認した．本研究の成果は，多様な価値観を変化させインタラクションを行う人間の認知過程のモデル化に寄与する．また，一律の設計で多様な行為主体に適したモデル構築を行える，汎用性の高いモデル構築手法の実現に寄与し得る．

キーワード：マルチエージェント, 多様な行為主体, 価値観, 秩序化, 強化学習

1. はじめに

人間はそれぞれ身体的特徴や価値観など異なるものを持っており，多様な行為主体が存在する複雑系の中で生活を行っている．多様な行為主体が存在する人間社会において，人間はルールや規制がトップダウンに定まっていない環境でもボトムアップにルールを形成し，秩序だった振舞いを行っていることがある．また，人間は複雑な環境で生きていくために価値観を状況に合わせて変化させ，その場その場に応じた行動をとっている．ここでの価値観とは行為の指標となるものであり，自動車運転を例にあげると「速く走る」「安全に走る」「燃費よく走る」などの価値観がある．このように人間は生活の中で課題を遂行するとき，いくつもある価値観を状況に合わせ，動的に変化させることで他者や環境と円滑なインタラクションをとっていると考えられる．このように価値観を動的に変化させ，行動を変化させる人間に対し，エージェントやロボットなどの機械はプログラムされた範囲で行動しているため人間が構築する秩序に適応できないといった問題がある．したがって人間社会で環境を共有する機械には道具として扱われるのではなく，人間や他エージェントとの社会的なインタラクションを想定した行動モデルが必要になってくると予測される．

人間が行う社会的なインタラクションをエージェントや機械に導入するとき，これまでは人間がトップダウンに決定したモデルを用いられていた．例えば，竹内ら (2000) や中西ら (2001) は決められた社会的応答に対してエージェントにモデルを付与し社会性を見出せることを示唆している [1][2]．これらはトップダウンに社会性を見出す方法であり，人間が想定した振舞いしか行わないため行動が限定される．トップダウンに見出した社会性は人間社会において特定の状況下では有効であるが，多様なシチュエーションのあ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．一方で，ボトムアップな設計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 [3]．

ボトムアップな設計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．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 [4][5]．また，大倉ら (2011) や山田ら (2018) は連続空間の協調課題に対してマルチエージェントの強化学習を適用し社会的なインタラクションが創発することを

示した [6][7].

これらのボトムアップな手法を用いた研究ではシンプルな課題に対して単一の報酬を与え，最適化が行われている．しかし，実社会において一つの課題を実行するとき，その内部には小さなタスクが連続的に存在し，これを解決しなければならない．移動する課題を想定すると，目的地に移動する過程で，合流する場合は周りの行為主体に速度を合わせたり，危険な地点では即座に対応できるように速度を落としたり，その状況に応じて求められる振舞いを変化する．このような連続的なタスクが存在する課題に対して単一な報酬を与えて強化学習を行っても，最適行動の獲得は難しい．この問題に対して，複数の報酬を与える手法や状態ごとに行動モデルを構築する手法が用いられている [8][9]．これらの手法はあらかじめ課題遂行時に報酬となる行為や分割する状態が明らかになっていれば設計することが可能であるが，最適な行動がわからない未知の課題や多様な行為主体がいる環境で想定できないインタラクションが存在する場合設計することができない．このような連続的に生じる課題や異なった特徴を持つ他者に対して，エージェントが人間のように価値観を動的に変化させ振る舞うことが可能であれば，高い適応性を持った行動モデルを示すことができる．

そこで，本研究では強化学習手法の課題である複雑な環境や連続的な問題への適応について，従来の手法のように単一の指標 (報酬) を与えるのではなく，人間が持つ多様な価値観のように自由度の高い指標を与えることで，学習の中で各エージェントがボトムアップに報酬系を構築し，多様な価値観を創発することができるか確認する．また，創発した価値観が何に起因しているのか確認し，人間が持つ多様な価値観のモデル化を行う．

本稿では最適化を行う課題として連続的なタスクと多様な行為主体を設計することができるマルチエージェントの移動課題を用いる．この移動課題を題材にして，環境や他者に応じて価値観を動的に変化させる振舞いが創発する可能性をシミュレーションによって確認することを目的とする．この移動課題は自動車の交通環境や混雑する駅や施設での人間の移動といった身近な問題を抽象化した課題として扱っており，創発された振舞いや価値観を現実空間に近い環境で確認できることが期待できるため強化学習課題に適応した．

本研究では，複雑な環境や連続的なタスクが存在する人間社会で多様な価値観を持ち，状況に応じて動的に価値観を変化させ対応している人間の認知メカニズムに着目し，価値観を動的に変化させる振舞いを強化学習によって実現することを試みた．

本研究の成果は，多様な価値観を動的に変化させインタラクションを行う人間の認知過程のモデル化に寄与する．また，多種多様なエージェントに対して一律の設計で，特徴を持つ各行為主体に適したモデル構築を行える，汎用性の高いモデル構築手法の実現に寄与し得る．

2. 移動課題

2.1 移動課題のルール

本研究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．

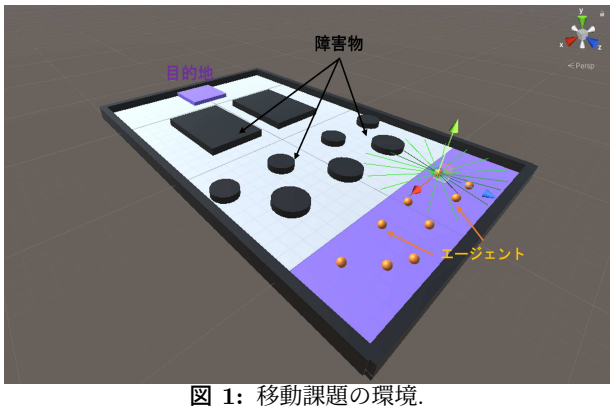


図 1: 移動課題の環境.

2.2 移動課題における多様な価値観

人間のような行為主体はそれぞれ身体的・物理的に特徴を持っており，その特徴によって行動可能範囲も異なってくる．行動可能範囲とは行動方策の枠組みであり，本研究で着目する価値観も枠組みが異なれば変化してくると考えられる．多種多様な行為主体が存在する環境であっても，人間は価値観を動的に変化させ秩序を保っている．そこで本研究では枠組みとなる特徴とそこから創発される価値観に着目し，モデル化を試みる．ここでは題材として扱う移動課題における行為主体の特徴と価値観の定義，また注目すべき振舞い

にを述べる．行為主体の特徴とは身体的・物理的特徴と内的表象を表す内部特性の二つがあげられるが，本研究では行動方策と価値観の関係性について着目するため，内部状態までの観測は行わない．本稿では「大きさ」「速度」「加速度」の 3 つの要素を用いて行為主体に身体的特徴を付与する．次に，移動課題において価値観としてあげられるものは「早さ」「安全性」「燃費」「快適性」などがある．人間が移動するときはこれらのような価値観を動的に変化させ行動していることがある．例えば,「早さ」に価値観をおいてスピードを出して移動していた人が，場所や他者との位置関係によって「安全性」に価値観を切り替えその場だけは周りをよく見て，スピードを落として移動するような行動をとることがある．このような価値観の変化に注目し分析を行う．次に観測する振舞いについて述べる．振舞いから価値観を評価するためには，課題を遂行する中で行為主体が行動の指標とするものを事前にあげ，その価値観に則する行動がどのようなものか定義しておく必要がある．題材として扱う移動課題は自動車が走行する状況や人間が徒歩で移動する状況を抽象化した課題であり，この移動課題における価値観は表 1 にまとめたものがあげられる．

表 1: 移動課題における価値観

価値観	内容
移動の速さ	移動速度が速いか
時間のロス	停止時間が短いか
安全性	他者と接触していないか
燃費	移動距離が短いか
快適性	急加減速をしているか

この表 1 のようにエージェントが実際にとった振舞いから現在の価値観を評価するため，振舞いに現れない行為主体の内部状態等については本稿では評価しないものとする．このような多様な価値観を変化させる振舞いをエージェント間の距離や速度といった行動データからモデル化することを目指す．これにより，エージェントが価値観を変化させる条件を示すことが可能になり，将来的には価値観を動的に変化させ，複雑な環境や多種多様な他者へ対応するエージェントモデルの構築に寄与すると考えられる．

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 [10] や Q 学習にニューラルネットワークを用いた DQN[11] やゲーム課題に用いられることがある PPO[12] などが存在する．どの学習法が本研究の題材である荷運び課題に適しているか比較によって調査した．本研究は将来的に実社会への応用を考えているため課題も 3 次元環境で行う．そのため，状態空間の情報を連続値で扱える手法が望ましい．また，様々な状態空間を考えていくため状態数が増加しても安定して学習を行える手法が望ましい．さらに，複数のエージェントで学習を行うようなマルチエージェントシミュレーションでは一度学習した振る舞いが他エージェントの学習により別のステップで最適ではなくなる可能性がある．したがって，マルチエージェントでの学習に対応可能な頑健性を持った手法が望ましい．これらの観点から，もっとも代表的な学習手法である Q 学習の状態空間は離散値を扱うため適していない．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 2 に示す．表 2 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 2: 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，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 [12][13]．PPO の特徴は方策関数を更新する際，その変化量が大きくならないようにするた

めにクリッピングを行うことで学習の安定化を行っている点である．図 2 に PPO のアルゴリズムのフローチャートを示す．方策の更新は式 1 を目的関数とした勾配法を用いる．クリッピングは式 1 中の clip 関数であり，式 2 に示す方策の変化の比が $1-\epsilon$ より小さい場合，および $1+\epsilon$ より大きい場合に変化量を一定の値にする処理である．式 1 の θ は方策パラメータ，$\hat{\mathbb{E}}_t$ は経験の期待値，$\hat{A}_t$ は Advantage の推定値，ϵ はハイパーパラメータを示している．

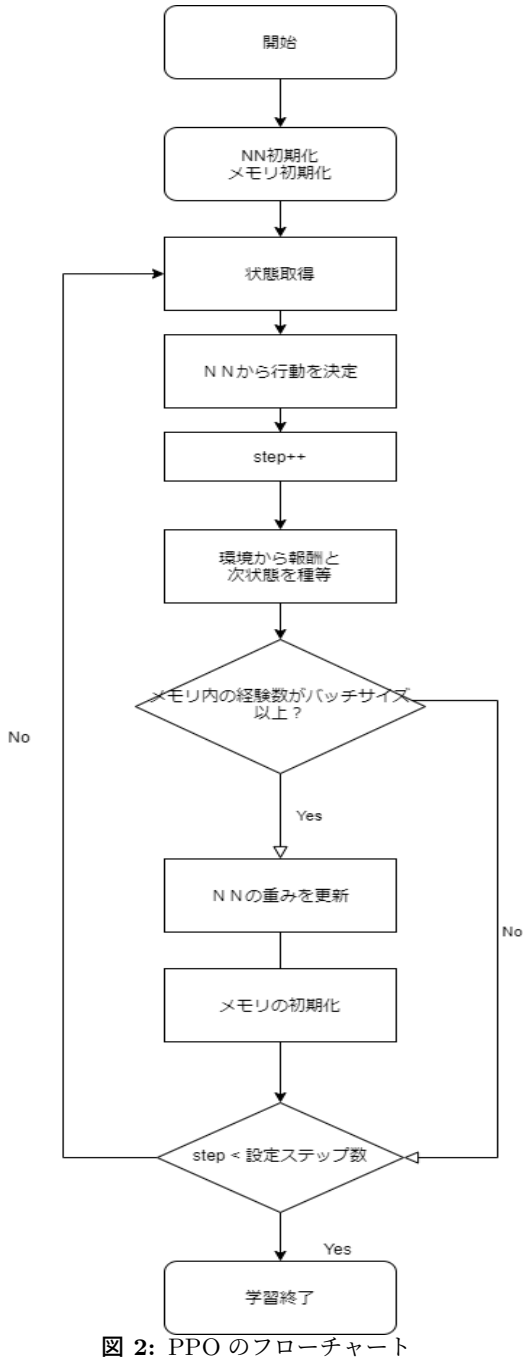


図 2: PPO のフローチャート

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

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

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，行動空間についての説明と PPO のハイパーパラメータの値を示す．報酬系は実験条件に関わってくるため次章に記述する．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 3 に示す．

表 3: 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))	50

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．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	512
隠れ層の数	3

4. 学習実験

4.1 実験環境

移動課題の環境は図 1 のように Unity で作成した．図 1 において，オレンジ色の球体がエージェント，黒い円柱と直方体は障害物，紫の立方体は目的地である．エージェントが課題を行うエリアは壁で囲まれており，エージェントは上部の紫のエリア内の位置にランダムに生成される．障害物と目的地は決められた位置に配置される．

エージェントは図 3 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．このような仮想環境を Unity 上に構築し，強化学習ライブラリである ML-Agents を用いて学習を行う．

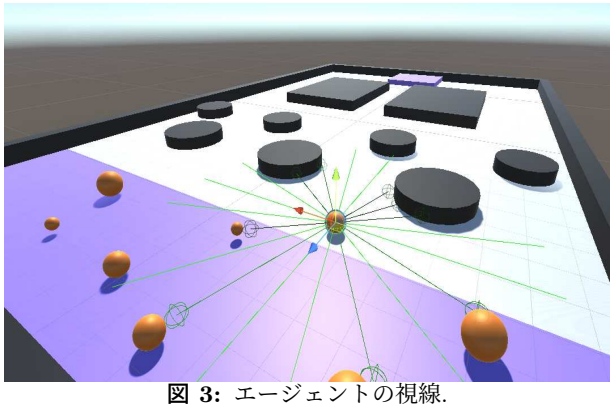


図 3: エージェントの視線.

4.2 実験条件と評価方法

本実験では移動課題を題材に，強化学習を用いて移動行動を最適化することで，どのような条件でエージェントが多様な価値観を構築するのか確認することを目的とする．今回の学習実験では構成論的なアプローチで条件を決定し，学習を行うことで多様な価値観の創発に各条件が寄与してくるのか議論を行う．

本稿では (1) 行動方策の枠組みとなる身体的特徴の有無と (2) モデル構築の指標となる多様な変数を用いた自由度の高い報酬設計の有無を条件に，強化学習を行うことでエージェントが多様な価値観を創発するか，またどの要因が多様な価値観の創発に寄与しているか確認する．

身体的特徴とは各エージェントに大きさ・速度・加速度の 3 変数をそれぞれ変化させ先天的に個性を与えるか否かである．多様な変数を用いた自由度の高い報酬設計の有無とは従来の強化学習で用いられている単一の行為やイベントに対して報酬を与える報酬系ではなく，課題を遂行する際に価値となる小さな行為 (強化学習において 1step 単位の行為) を複数あげ，それらに対し小さな報酬を与えた自由度の高い報酬系である．身体的特徴を与えることで方策の枠組みが多様化されるため，創発される方策も多様なものになると考えられる．また，自由度の高い報酬系を与えるとエージェントは学習を行う過程で自身の報酬系をボトムアップに構築でき，状況に応じた多様な価値観を獲得可能であると考えられる．以上のことから本実験は表 5 の 4 条件で強化学習を行う．

表 5: 実験条件

条件名	身体的特徴 (Charcter)	多様な変数の報酬系 (Diverse Reward)
CDRL	○	○
CL	○	×
DRL	×	○
L	×	×

CDRL(Charcter and Diverse Reward Learning) 条件は, 各エージェントに大きさ，速度，加速度の 3 変数をそれぞれ変化させ身体的特徴（個性）を与える．さらに，多様な変数を用いて報酬系を設計し，最適化の指標に自由度を与えた状況下で学習させる条件である．CL(Charcter Learning) 条件は, 身体的特徴は与えるが，最適化の指標に自由度を与えない状況下で学習させる条件である．DRL(Diverse Reward Learning) 条件は, 身体的特徴は与えずに最適化の指標に自由度を与えた状況下で学習させる条件である．L(Learning) 条件は, 身体的特徴も最適化の指標に自由度も与えない状況下で学習させる条件である．以上の 4 条件で学習シミュレーションを行う. そして,CDRL 条件と CL 条件,DRL 条件と L 条件を比較することで最適の指標に自由度を持たせることが創発する価値観に寄与するか検討し,CDRL 条件と DRL 条件,CL 条件と L 条件を比較することでエージェントに身体的特徴を持たせることがボトムアップに構築する価値観に寄与するか検討する．

それぞれの条件において 9 体のエージェントを用意し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行った．各条件の報酬設計は表 6 のように行った．報酬系に自由度を与えない状況下で学習させる条件の CL 条件と L 条件は，中間地点・目的地までの移動にのみ報酬を与え，時間切れと他者接触に対しペナルティを与える従来の強化学習手法と同様の簡素な報

表 6: 条件別の報酬設計

イベント	報酬・ペナルティ	値	CDRL，DRL 条件	CL，L 条件
目的地まで移動	報酬	0.8	○	○
中間地点まで移動	報酬	0.2	○	○
時間経過	ペナルティ(毎ステップ)	0.0001	○	×
オブジェクトとの接触	報酬・ペナルティ(毎ステップ)	0.0001	○	×
停止時間	報酬・ペナルティ(毎ステップ)	0.0001	○	×
急加減速	報酬・ペナルティ(毎ステップ)	0.0001	○	×
移動距離	報酬・ペナルティ(毎ステップ)	0.0001	○	×
他者と接触	ペナルティ	0.7	○	○
時間切れ	ペナルティ	0.7	○	○

酬系で学習を行う．一方で報酬系に自由度を与える状況下で学習させる条件の CDRL 条件と DRL 条件は，簡素な報酬系に加え，表 6 のように小さなイベントに対して報酬・ペナルティを与え学習を行う．

4 条件とも 50Mstep の学習を行い，条件別の学習過程とエージェントが獲得した行動モデルを比較することで評価を行う．学習過程では学習が収束するまでの時間を比較する．エージェントが獲得した行動モデルは，学習後のエージェントがとった行動からエージェントがどうような価値観を持っているのか定量化し，価値観が状況によって変化しているか分析する．定量化は表 1 に示している価値観に対応する行動がどの程度行われているかで評価する．

4.3 学習結果と考察

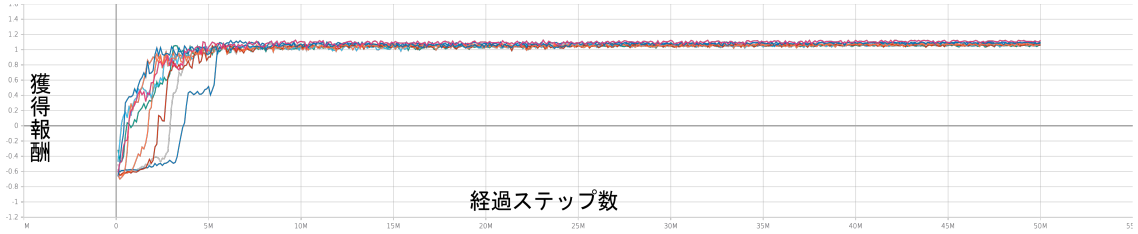
4.4 多様な行為主体と報酬系の変化による学習進度の違い

PPO を用いて各条件で 50M ステップ学習した結果を図 4 に示す．なお，図 4 にはエージェント 9 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．図 4 はすべてのエージェントの学習経過を示しており，学習進度について条件ごとに比較すると CDRL 条件で学習を行った 9 体のエージェントが最も早く収束している一方で，CL 条件のエージェント全体の学習進度は最も遅い．CL 条件は報酬系を細かく多様な変数を用いて設計しておらず，単調な報酬設計のもと学習した条件であるため目的地まで到達する行為の最大化を行う条件である．一方で CDRL 条件は多様な変数を用いて報酬系を設計しているため，目的地まで移動する中でも細かい行為（接触しない移動や迅速な移動など）に対し最大化を行う条件である．このように報酬系に自由度を持たせているため CDRL 条件の学習は探索範囲が広がり学習効率が悪くなると予測していたが，CDRL 条件は他 3 条件に比べ，報酬の最大化が素早く行われている．この要因としては，細かい報酬やペナルティを設けることで初期の探索的な学習を行うときに目的地まで移動せずとも適した行為を行えば報酬を得れるため，初期のランダム行動での学習がはかどったことが考えられる．また，9 体のエージェントが報酬を最大化させるまでの時間差を見ると，CDRL 条件は最も時間差がなく同時に学習が行われている．これは複数エージェントが存在する環境で同時に学習することが可能であることを示しており，さらに多種多様な環境下での学習においても有効であることを示唆した．

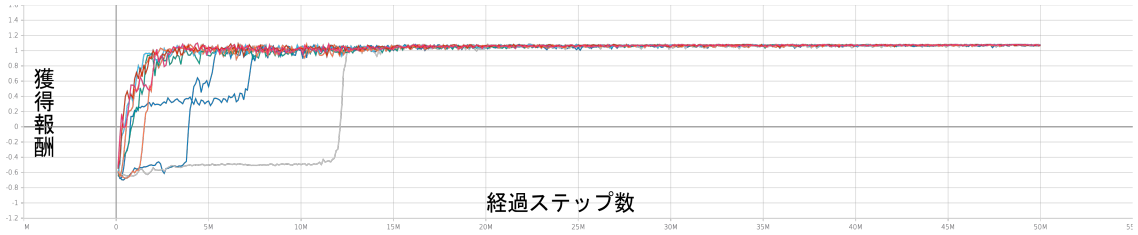
4.5 自由度の高い指標が構築する秩序

次に，エージェントが学習によって獲得した行動方策からエージェントが創発した価値観について考察する．行動方策といっても位置や速度，加速度，他者の位置など様々な要素から分析が必要であるが本稿ではエージェントの速度と位置関係から価値観について言及する．図 5 には学習後の行動モデルを搭載したエージェントを同環境で 30 試行した際の速度分布を環境上にプロットしている．なお，分布図に 9 体の軌跡を示すと乱雑になるため 1 体のエージェントの行動軌跡を示している．

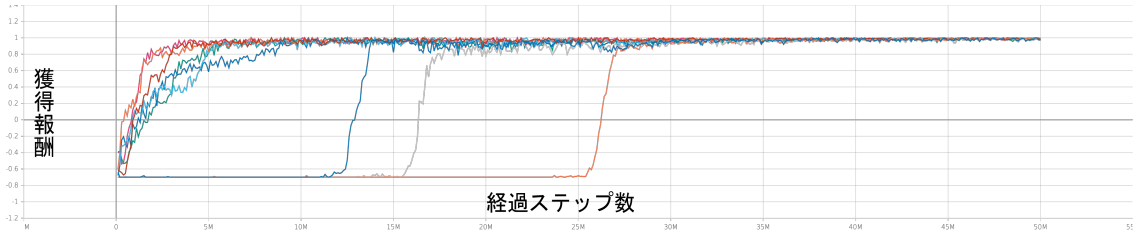
図 5 を見るとそれぞれの条件において最大速度で走行している場所と速度を落として走行している場所が存在する．このように最大速度で走行している場所はエージェントが移動の速さに価値観をおいて行動していると捉えることができ，速度を落として走行してい



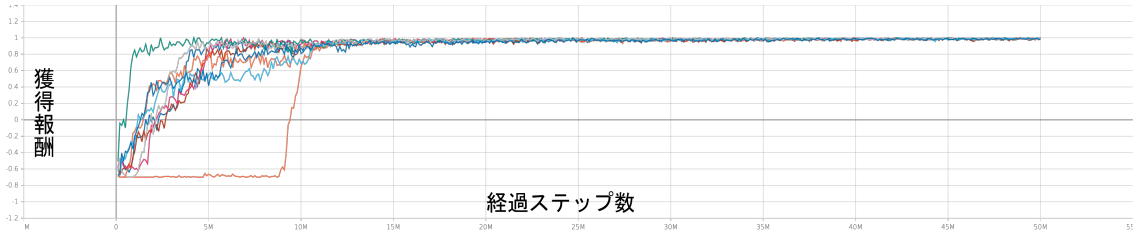
(a) CDRL 条件



(b) DRL 条件



(c) CL 条件



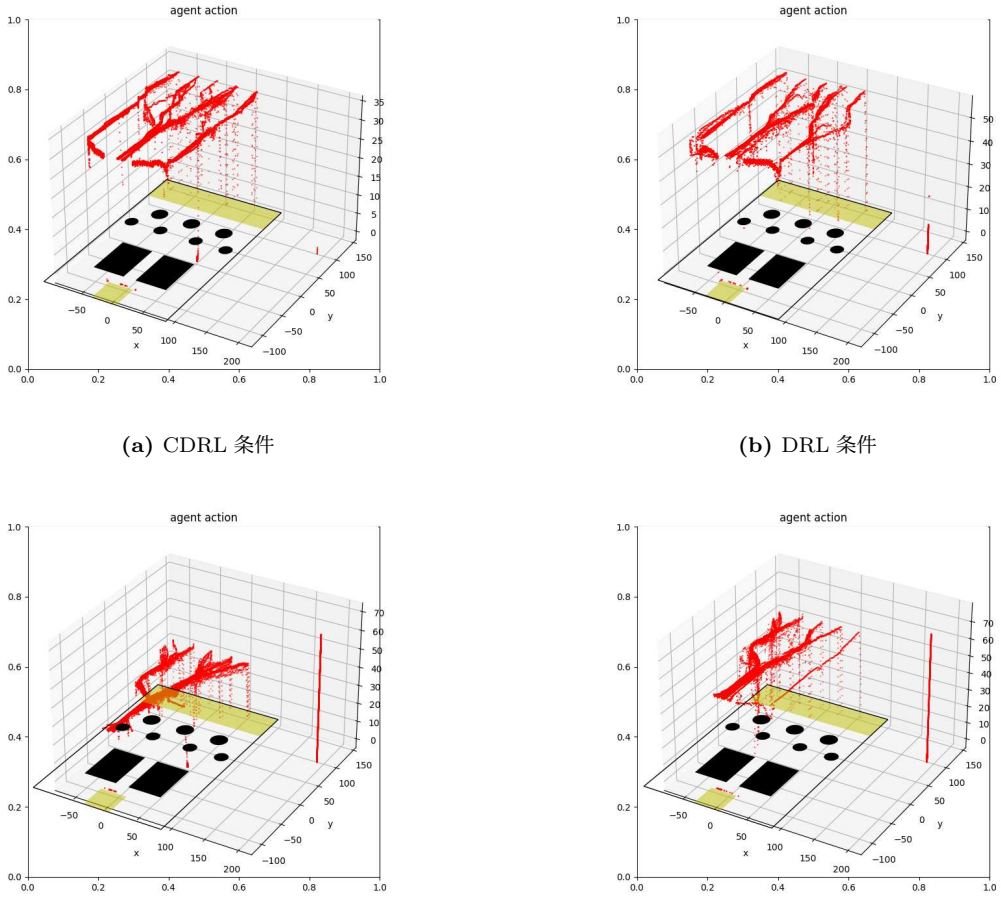
(d) L 条件

図 4: 学習の結果 4 条件でエージェントが獲得した報酬の推移

る場所はそれ以外の価値観を重視して行動していると考えることができる．この結果は速度の観点から価値観を捉えると，価値観を変化させ行動するエージェントの構築はすべての条件において可能であると示唆している．次に，これらの速度変化が環境との位置関係に起因しているのか調査した．図 6 に 9 体のエージェントが減速して走行した場所をプロットしたもの示す．

図 6 を見ると L 条件と CL 条件では中央の狭路前で減速が良く発生している．これは多くのエージェントが最短のルートである狭路を走行するという方策を構築した結果，狭路にエージェントが密集し，減速が発生したと考えられる．CDRL 条件と DRL 条件でも狭路前に減速は発生しているが CL 条件，L 条件と比較すると狭い範囲でしか発生していない．自由度の高い報酬系のもと学習したことにより最短のルートを走行する以外の価値を見出し，別のルートを走行する方策を学習した結果，密集を抑えることができたと考えられる．この結果は，多様な変数を用いた報酬系のもとでの学習は単調な報酬系のもとでの学習に比べ，方策が一意に決まらず，多様な価値観を創発可能であることを示唆した．

また，環境に創発した秩序をエージェントの速度の観点から捉えると，CL 条件・L 条件では広い範囲で減速が求められ，特に中央にある狭路付近ではより減速が必要とされ，速度規制が強い秩序が形成された．一方で，CDRL 条件・DRL 条件において減速が求められる範囲は狭く，自身の最大速度で走行可能な範囲が広く設けられるおり，環境において速度規制が弱く，自由度の高い秩序が創発された．ここでの秩序とはあらかじめ与えたトップダウンの規則ではなくエージェ



(a) CDRL 条件 **(b)** DRL 条件

(c) CL 条件 **(d)** L 条件

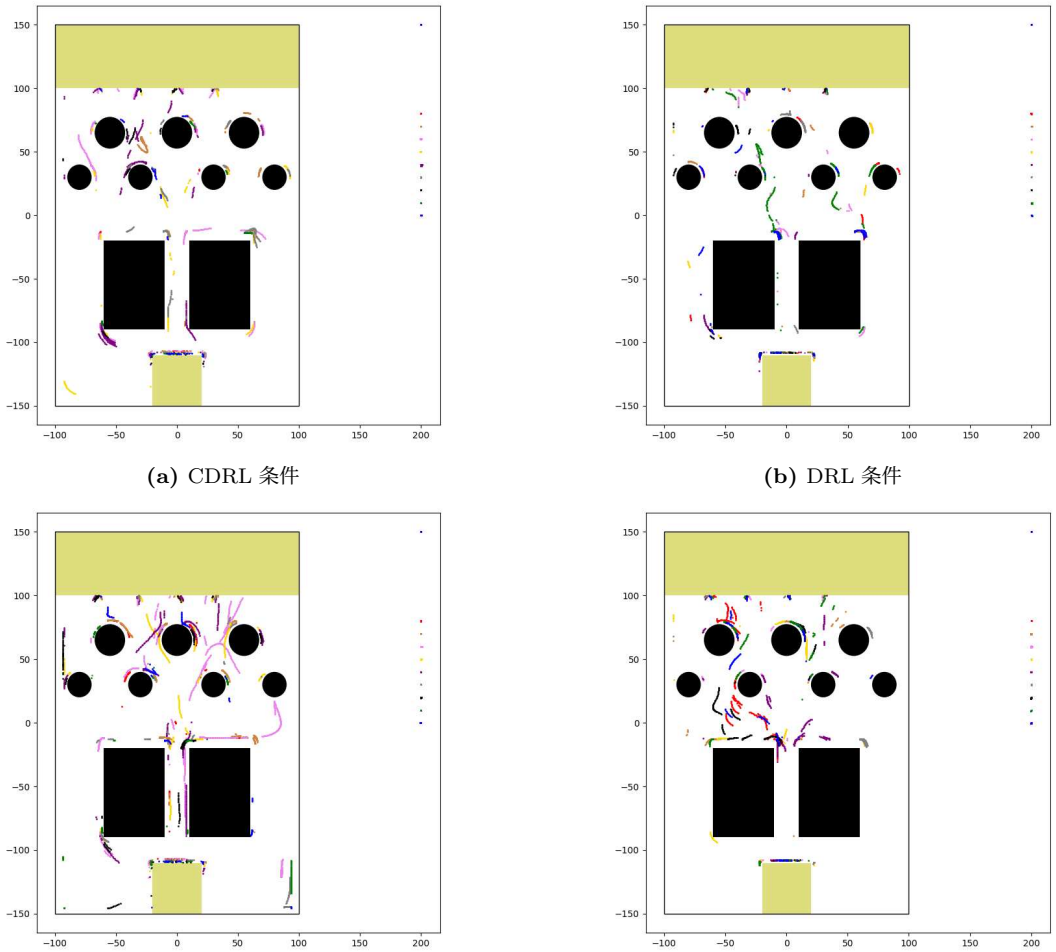
図 5: 学習の結果獲得したエージェントの行動 (速度分布)

ント群が試行錯誤を行う中でボトムアップに構築した規則と定義している．この結果は環境に創発した秩序の違いを示しており，単調な報酬系のもと学習を行う CL 条件・L 条件は単一の指標に向けて学習を行うため，身体的特徴が異なる行為主体が存在する場合では最適化が一意に定まり，似通った方策が構築され，規制が強い秩序となったと考えられる．一方で CDRL 条件・DRL 条件では多様な行為主体が存在しても，各行為主体が自身に合った最適化を行えるためエージェントごとに異なる方策が構築され，自由度が高い秩序が形成されたと考えられる．

5. おわりに

本稿では多様な価値観のもと行動する人間の振舞いをエージェントに実現するための足掛かりとして，強化学習によってボトムアップに行動モデルを構築するさい，多様な変数を用いた報酬系の有無と身体的特徴の有無が創発する価値観に変化を産むか検証した．学習の結果，学習進度の観点では多種多様な行為主体が存在する場合，多様な変数を用いた報酬系を扱うことで効率よく学習可能なことが示された．また，創発した秩序に着目すると，多様な変数を用いた報酬系のもと学習することでエージェントの個性を最大限生かした自由度の高い秩序が創発されることが示された．

本研究は，自動車での交通環境上の移動や人間の施設内の移動を抽象化した移動課題を題材として，連続的なタスクが起こりえる状況を想定し学習実験を行い，状況に応じて価値観の変化が起こっているのか分析した．実環境に近い連続空間においてシミュレーションを行い，各エージェントの振舞いやインタラクションを観察できるため，複雑環境の中で人間のように状況に応じて価値観を変化させ行動するエージェントのインタラクションモデルの構築に寄与し得る．また，本研究では行為主体が集団内で行動する状況において秩序形成の過程や条件について言及しており，人



(a) CDRL 条件 **(b)** DRL 条件

(c) CL 条件 **(d)** L 条件

図 6: 学習の結果獲得したエージェントの行動 (減速時の速度分布)

間が他者と環境を共有する中で秩序を形成するメカニズムの解明に寄与する．

今後の課題として，まずは他者とのインタラクションを考慮した秩序の評価があげられる．本稿ではエージェントが創発した秩序について，速度と静的な環境との位置関係から評価を行った．そのため，動的な環境（他エージェント）の存在が与える秩序への影響については言及できない．人間が他者と環境を共有し，課題を行う場合，少なからず他者の存在は考慮していると考えられる．例えば，一方の道に大きな行為主体がいるからもう一方の道を選択するといった振舞いは良く起こる．このように他者の情報は行動する際の価値観に影響を与えていると考えられるため，他者との位置関係や能力差を含め，創発した秩序を評価する必要がある．また，今回は身体的特徴の有無と多様な変数を用いた報酬系の有無の組み合わせが創発する動的価値観に変化を及ぼすか確認した結果，多様な変数を用いた報酬系の有無によりマップの減速ポイントの範囲に差が見られた．しかし，減速は近くに存在する他者を避ける際に発生しており，他者との位置関係が行動に強く依存していると考えられる．これは他者を広範囲で認識していたことが要因としてあげられるため，今後は多種多様な行為主体がいる環境で多様な変数を用いた報酬系のもと，他者の認識範囲を条件に実験を行い，創発する秩序に変化が生じるのか調査する予定である．

文献

[1] 竹内勇剛, 片桐恭弘 (2000).“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.
[2] 中西英之, 石田亨 (2001).“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.
[3] 山村雅幸, 宮崎和光, 小林重信 (1995).“エージェントの学習”, 人工知能学会誌, Vol.10,No.5,pp.683-689.
[4] Nagata,Y.,Ishikawa,S.,Omori,T.& Morikawa,K.(2007).“Computational model of coop-

erative behavior: Adaptive regulation of goals and behavior”,Proceedings of the European Cognitive Science Conference,pp.202-207.
[5] 保田俊行, 大倉和博 (2013).“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.
[6] 大倉和博, 保田俊行, 松村嘉之 (2011).“構造進化型人工神経回路網による Swarm Robotics のための適応的協調行動の生成”, 日本機械学会論文集,Vol.77,No.775,pp.399-412.
[7] 山田和明, 保田俊行, 大倉和博 (2018).“マルチロボットシステムのための状態空間表現を適応的に切り替える強化学習”, 日本機械学会論文集,Vol.84,No.862,pp.17-288.
[8] 内部英治, 銅谷賢治 (2004).“複数報酬のもとでの階層強化学習”, 日本ロボット学会誌,Vol.22,No.1,pp.120-129.
[9] Patrick,K. & Rudolf,M.(2019).“Simulated Autonomous Driving in a Realistic Driving Environment using Deep Reinforcement Learning and a Deterministic Finite State Machine”,Proceedings of the 2nd International Conference on Applications of Intelligent Systems，pp.1-6.
[10] Amy,G. & Keith,H.(2003).“Correlated-Q Learning”,ICML,pp.84-89.
[11] Yingfeng,C.,Shaoqing,Y.,Hai,W.,Chenglong,T., & Long,C.(2021).“A Decision Control Method for Autonomous Driving Based on Multi-Task Reinforcement Learning”,IEEE Access2021, pp.154553-154562.
[12] JunLAI, Xi-liang,C., & Xuezhen,Z.(2019).“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.
[13] Bøhn, E., Coates, E. M., Moe, S., & Johansen, T. A.(2019).“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.