他者の振る舞いに応じた秩序形成のための インタラクションの検討

Investigation of Interaction for Order Formation According to the Behavior of Others

原田 雄大 *　　竹内 勇剛
Yudai Harada　　Yugo Takeuchi

静岡大学
Shizuoka University

Abstract: 他者と環境を共有する状況でエージェント同士のインタラクションをトップダウン的にモデル化してしまうと，限定した状況でしか有効でなく，多様な状況が想定される場合には汎用性に欠ける．一方で，ボトムアップ的な手法によるインタラクションのモデル化は多様な状況に適応した柔軟性を有する可能性があるが，エージェントが多数であったり多種であったりすると指数乗的に複雑さが増してしまう．そこで本研究では，他のエージェントの振る舞いを環境変化の一部として取り扱うことによって各エージェントは環境とのインタラクションを行うことで，ボトムアップ的な手法を用いても計算量を抑えつつ高い環境適応性を得られると考え，強化学習を通した 2 体のエージェント間インタラクションを観察する．本稿では題材として個での振る舞いと全体の構造としての動きの両方を観察することができる荷運び課題を用いて，他者情報の参照範囲を条件にして実験を行った．その結果，各エージェントは他のエージェントと直接インタラクションを行わなくてもエージェント同士が組織化された秩序を創発していることが確認できた．ここでの秩序とはボトムアップに創発した，あらかじめ与えていないルールと定義している．この結果は，エージェントは環境に立脚した限定的な情報しか得られなくても，環境を介して他のエージェントと適応的なインタラクションを成り立たせることができる可能性を示唆しており，エージェント間インタラクションに関するモデル構築の新たなデザインに寄与することが期待される．

1　はじめに

昨今のハードウェアやソフトウェア技術は急激に発展しており機械が知能を持つ人工知能や人の代わりとなり機械が運転を行う自動運転など様々な機械が生み出されている．自動運転車や対話エージェントなど日常的にあるものの自動化が研究，開発されていく中で，人間がロボットやエージェントと協調しながら同環境を共有することが将来訪れると考えられる．このような状況で人間は他者に応じて振る舞いを変更し，秩序を形成していると考えられる．ここでの秩序とはあらかじめ存在するルールとは別のボトムアップに創発した規則と定義している．倉本 (2012) は，エージェントに個性を付与することで，ユーザのエージェントに対する印象が変化することを実験により示しており，岡 (2018) はインタラクション持続の要因として個人の特性に着目し，インタラクション実験により個人特性がインタラクション持続に影響を与えることを示唆していた [1][2]．つまり，人間が他者とインタラクションする際，相手の個性や属性に応じて行動を変化させ，協調的に振舞い，秩序の形成を図っていると考えられる．ここでの個性とは他者との弁別性であり，同種の機能を持つ行動主体にある要素において異なる機能を持たせることである．個性にも身体的・物理的な特徴を表すものや内部状態に付与するものが考えられる．自動車を運転する状況を考えたとき車の大きさや速さは身体的・物理的特徴として個性として挙げられるが，運転者の運転特性のような内部的特徴も個性として挙げられる [3]．本研究では個性として身体的・物理的特徴に着目し，他者とのインタラクションに変数として取り扱う．

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 (2000) や中西 (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

*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

ことを示唆している [4][5]．これらはトップダウンに社会性を見出す方法であり，人によって決められた振る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，人間の秩序だった振る舞いは状況によって，既存のルールから外れた行動をとることがある．このように，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．

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に適した振る舞いを獲得できるため，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(2007) や保田 (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，集団内での協調が創発することを示している [6][7]．また，大倉（2011）や山田（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協調が創発することを示した [8][9]．これらのボトムアップな手法を用いた研究では各エージェントのインタラクションを想定し，学習を行っている．しかし，エージェントが多数であったり，多種であったりすると指数乗的に複雑さが増し，計算量が増えたり，モデル自体が複雑になることが考えられる．他者を環境の一部として扱い，情報を限定した状況でも秩序をボトムアップに創発し他者と環境を共有できれば，エージェントが多数であったり，多種であっても計算量を抑えつつ高い環境適応性を持った汎用的なモデルを示すことができる．そこで本研究では，個性を持った各行動主体の個別のインタラクションを考えるのではなく，他の行動主体を環境の一部としてボトムアップに行動し，高い環境適応性を備えた秩序を創発するかに着目する．

本稿では題材として個での振る舞いと全体の構造としての動きの両方を観察することができる荷運び課題を用いる．この荷運び課題は自動車の交通環境や人間に対して混雑する駅や施設でのナビゲーション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，複数体のエージェントに同時に行わせることで他者と共有し，行動する環境を構築できる．また，属性を付与した複数体での荷運び課題をボトムアップなアプローチ法である強化学習によって，各エージェントの振る舞いを最適化することで他者の属性に応じた振る舞いを獲得できると考えられる．そこで今回の実験ではエージェントの大きさという要素を個性として与えた複数体のエージェントを同環境に混在させた中で課題の学習を行った．実験は他者を環境の一部として考え，情報を限定したときに高い環境適応性を備えた秩序を創発するか検証するために他者情報を認識できる範囲を条件にして比較を行った．その結果，他者の情報を限定して学習を行った条件でもエージェント同士が組織化された秩序を創発していることが確認できた．

本研究の成果は，ボトムアップな手法でも他者を環境の一部として捉え，情報を限定することで計算量とモデルの複雑さを抑えつつ，汎用性の高いモデルを実現可能かを議論することに繋がり，多種多様な行動主体が存在する人間社会の中で，他者と環境を共有するインタラクション構造の解明に寄与できる．

2　荷運び課題

本章では実験で取り扱う課題の妥当性を示すために課題の持つ機能性について記述する．

2.1　課題のルール

荷運び課題とは図 1 のようにフィールド内にスタート地点とゴール地点（黄色い地点）を決め，フィールド内の制約を従い，他者と環境を共有しながらスタート地点とゴール地点を行き来して指定数の荷物を運ぶ課題である．図 1 のように黒い障害物を配置してある環境上で行動主体であるエージェント（オレンジの球体）が走行する．エージェントは環境に与えられた制約の範囲内で行動する．指定数の荷物を運ぶと課題成功で終了するが，課題中に他エージェントとの接触や時間切れなど制約を違反すると課題失敗で終了する．

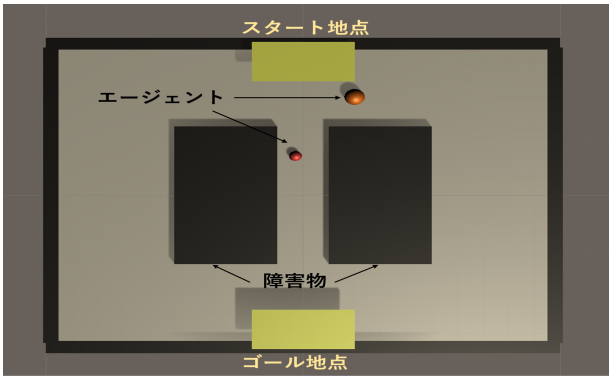

図 1: 実験環境.

2.2　荷運び課題における振る舞い

本研究ではエージェントの振る舞いを観測することで考察を行う．ここでは，荷運び課題における注目すべき振る舞いを述べる．

荷運び課題の一般的な例として自動車が走行する交通環境がある．自動車はそれぞれが別の目的地を目指し，走行しているが，目的地に到達するために協調的な行動をとることが良くある．パーソンズ (1989) は目的や目標が統一されているとき，そこに秩序が形成されると述べている [10]．本研究の題材である荷運び課題においても荷物を運ぶという共通の目標があるため協調的な振る舞いを行い，秩序が形成されると考えられる．秩序とは他者と環境を共有するときに，あらかじめトップダウンに与えたルールではなく環境や他者とのインタラクションの中でボトムアップに創発されたルールと定義している．

荷運び課題における個体の振る舞いとしては道を譲ったり，遠回りをするといった協調的な振る舞いや個の利益のみを考えて最短距離だけを走行するといった利己的な振る舞いが挙げられる．本研究では荷運び課題をマルチエージェントでボトムアップに学習することにより，全体としての秩序が形成され，協調的な振る舞いがが創発されるのか実験により評価する．本課題では以下のように振る舞いを分類し，評価を行う．

無秩序な振る舞い： あらかじめ与えた制約以外に全体としてルールが観測できない振る舞い

秩序だった振る舞い： あらかじめ与えた制約以外に全体としてルールが観測できる振る舞い

また，我々人間が秩序形成を図るとき，他者の個性や属性に応じて行動を決定していると考えられる．倉本 (2012) や岡 (2018) は行動主体の個性に着目し，個性をパラメータとして考えることでインタラクションに変化があることを示している [1][2]．荷運び課題では環境に制約を設け，エージェントには身体的な特徴を与えるため，エージェントは環境と他者の二つの要素とインタラクションが行われる．本課題ではエージェントの行動モデルの入力として環境情報と他者情報を変数として扱うのでエージェントはエージェントと環境の間でインタラクションが行われ，環境を共有する他者とは図 2 のように個別でインタラクションするのではなく図 3 のように環境の一部としてインタラクションが行われる．

本研究ではこの荷運び課題をマルチエージェントに行わせ，学習を行うことで複数体で環境を共有しタスクをこなす振る舞いをボトムアップな手法により最適化する．本課題では身体的特徴である個体のサイズを個性としてエージェントに付与しており，この個性情報を他者のパラメータとして参照させる．個性を持つ他者との一対一のインタラクションではなく，他者が存在する環境とのインタラクションを行うことで，ボトムアップに秩序が形成され他者に応じた振る舞いを獲得すると考えられる．したがって，この荷運び課題は各個体が固有に有する性質に対して，それらを他者とのインタラクションにパラメータとして利用することで，各個体の振る舞いや全体の秩序にどのような変化をもたらすか観察することができる課題であると考えられる．

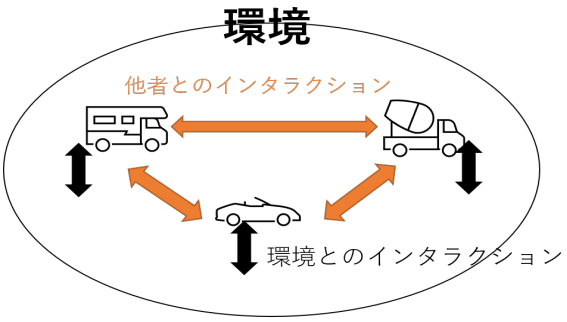

図 2: 個々のインタラクション．

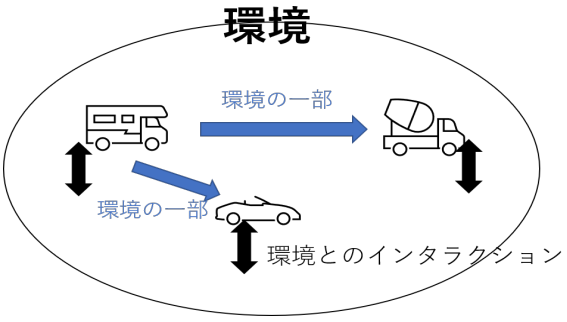

図 3: 環境とのインタラクション．

荷運び課題は身体的特徴を個性として持つ行動主体が，物理的制約のある環境において他者との共有が見られる課題である．この荷運び課題を一般化した時，自動車の交通環境でのナビゲーションや大規模な被災時やパンデミック時の患者の病院割り当てのような行動主体に個性があり，環境に制約がある状況での全体の秩序化に応用できるモデルを示すことができると考えられる．

3　強化学習手法

本章ではモデルを構築する手法として用いる強化学習について記述する．

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 [11] や Q 学習にニューラルネットワークを用いた DQN[12] やゲーム課題に用いられることがある PPO[13] 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 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荷運び課題は PPO が適していると結論付けた．次節ではＰＰＯのアルゴリズムについて記述する．

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，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 [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 4 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．次節では強化学習における各種パラメータについて記述する．

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

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

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．本節では荷運び課題でのエージェントの状態空間，行動空間，報酬についての説明と PPO のハイパーパラメータの値を示す．荷運び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に加え，他エージェントの位置や特徴（サイズ）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
スタート地点の絶対位置座標	2
ゴール地点の絶対位置座標	2
他エージェントの情報（位置情報とサイズ）	エージェント数*3

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．報酬に関しては，荷運び課題であるため荷物を持った時ときや荷物を運んだ時の行動に対して随時報酬が与えられるように設定した．また，報酬はステップが経過するにつれ減少するようにしており，早く課題を終了させたほうが高い報酬を得れるように設計している．ペナルティとしては時間

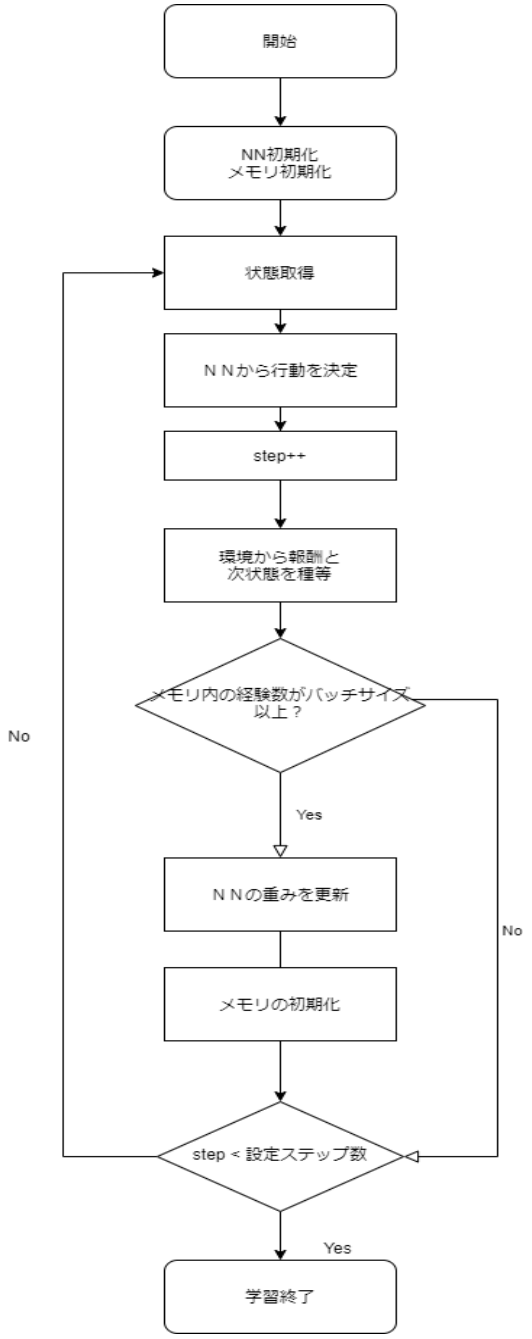

図 4: PPO のフローチャート

内に到着しなかったときや接触を起こした時に与えるよう設定した．表 3 に設定した報酬をまとめる．また，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 3: エージェントの報酬

内容	値
荷物を一回運ぶ	+0.2-(0.00005*step 数)
荷物を持つ	+0.2-(0.00005*step 数)
時間内に到達できない	-0.5
接触	-0.5

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	20240
方策変化量の閾値 ϵ	0.2
エントロピー正規化率 β	0.05
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

本章では行った学習実験について記述する．

4.1　実験方法

本研究では荷運び課題を題材にして，ボトムアップな手法である強化学習によって，他者の情報（位置情報と特徴情報）の参照範囲を変化させた状況において協調的な振る舞いを獲得し，秩序を形成することを検証する．本学習実験では仮想環境内に図 1 に示したようなフィールドを構築した．本実験では他者と環境を共有する状況で他者情報を限定したときに秩序の形成を行うのか検証することを目的としているため，エージェントは複数体用意する必要がある．Uwano ら (2019) はエージェントが 2 体でも協調な振る舞いを獲得し，ボトムアップに規則を創発できることを示唆しているため [14]，今回の実験ではエージェントを最小の複数体である 2 体で学習を行った．エージェントには個性とし

て身体的特徴であるサイズに差異を与え，小さなエージェントと大きなエージェントを用意した．また環境には障害物を図 1 のように配置し，中央の狭い道路と両サイドの広い道路を設け，道幅に制約を与えた．この環境のなかでそれぞれのエージェントが強化学習で荷運び課題の最適化を行う．

4.2 実験環境

実験は荷運び課題を行う環境を unity を用いて 3 次元の仮想環境内に構築し，学習を行う．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 10 Pro (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.2019.4.17f1)
強化学習ライブラリ	ML-Agents (release9)

4.3 実験条件と評価方法

本研究は表 6 に示す 4 条件を比較し分析を行う．表 6 に示すようにエージェントが他者の情報を参照できる範囲を条件に学習を行う．この荷運び課題においての他者を参照できる範囲とはエージェントからどのくらいまで離れた他エージェントの情報を状態変数として参照するのかというものである．

表 6: 学習条件

条件名	内容
C0	他者情報を参照できる範囲を 0 %
C25	他者情報を参照できる範囲を 25 %
C50	他者情報を参照できる範囲を 50 %
C100	他者情報を参照できる範囲を 100 %

条件 C0 は他者情報を参照できる範囲を 0 %としており，他者の情報を参照しない条件である．条件 C100 は他者情報を参照できる範囲を 100 %にしており，同環境にいるすべての他者の情報を参照する条件である．条件 C25 と条件 C50 は他者情報を参照できる範囲を 25 %，50 %に限定しており，環境全体を 100 %として，その 25 %，50 %の範囲のみ参照する条件である．この 2 つの条件では範囲外にいるエージェントの情報は参照せず，範囲内にいるときのみ情報を取得できるようにしている．このように他者情報を参照させる範囲を変化させ学習を行うことで，条件ごとにどのような秩序が形成され，また秩序の形成する過程に変化があるのかといった観点で分析を行っていく．

評価は，PPO によって獲得したエージェントの振る舞いを観察することで行う．評価を行うにあたり以下のように振る舞いを定義した．振る舞いの観察では学習途中のエージェントの動きと学習後の動きを比較し，秩序だった振る舞いを獲得しているか評価する．次に条件別の学習後の振る舞いを比較し，それぞれの秩序に違いがあるのか検証する．

無秩序な振る舞い： あらかじめ与えた制約以外に全体としてルールが観測できない振る舞い

秩序だった振る舞い： あらかじめ与えた制約以外に全体としてルールが観測できる振る舞い

実験では，まず強化学習により条件ごとに秩序が形成されたか比較する．秩序の形成が確認できた場合，他者情報の参照を限定し，計算量を抑えた時の学習時間について比較し，計算量を抑えられているか確認する．しかし，一回の学習データだと偏りがでることも考えられるので，同条件でもう一度学習を行い，二回分の学習データから評価を行った．学習時間を評価するにあたり，学習収束の条件を学習の獲得報酬遷移をもとに指数移動平均 (tensorboard の smoothing 機能) の係数を 0.9 にし，全エージェントの獲得報酬が 0.6 を超えたとき（一定以上の報酬を安定して獲得できるようになったとき）と定義する．この収束条件をもとに学習収束までに経過したステップ数を比較し，各実験条件の計算量に差異があるか検証する．

4.4 実験結果：秩序形成の比較

本節では条件別に学習を行い，エージェントが獲得した行動モデルを示し，創発された秩序について比較する．

4.4.1 C0

条件 C0 の学習後のエージェントのとった軌跡をプロットしたものを図 5 に示す．赤のプロットが小さなエージェントの軌跡を示しており，青のプロットは大きなエージェントのプロットを示している．条件 C0 の学習後の振る舞いは小さなエージェントが右側の広い道を走行し，大きなエージェントが真ん中の狭い道路を走行する結果となった．

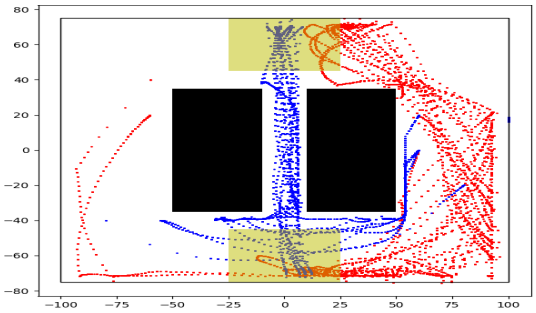

図 5: C0-振る舞い.

4.4.2 C25

条件 C25 の学習後のエージェントのとった軌跡をプロットしたものを図 6 に示す．条件 C25 の学習後の振る舞いは小さなエージェントが真ん中の狭い道路を走行し，大きなエージェントが右側の広い道を走行する結果となった．

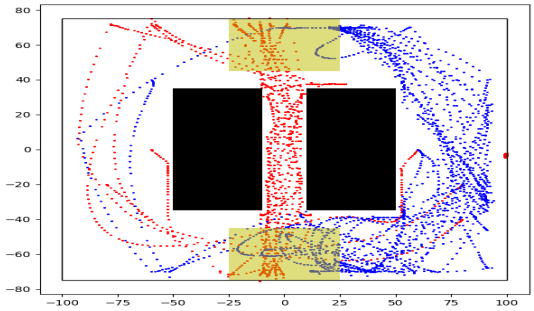

図 6: C25-振る舞い.

4.4.3 C50

条件 C50 の学習後のエージェントのとった軌跡をプロットしたものを図 7 に示す．条件 C50 の学習後の振る舞いは小さなエージェントが外側の広い道路を回るように走行し，大きなエージェントが真ん中の狭い道を走行する結果となった．

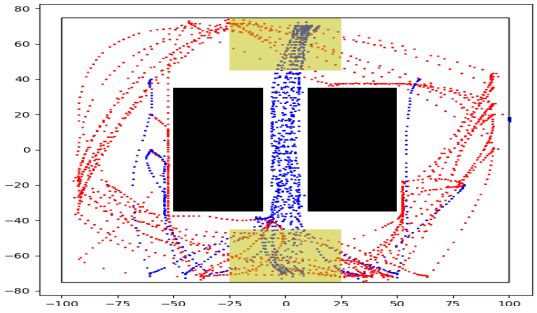

図 7: C50-振る舞い.

4.4.4 C100

条件 C100 の学習後のエージェントのとった軌跡をプロットしたものを図 8 に示す．条件 C100 の学習後の振る舞いは小さなエージェントが真ん中の狭い道を走行し，大きなエージェントが外側の広い道路を回るように走行する結果となった．

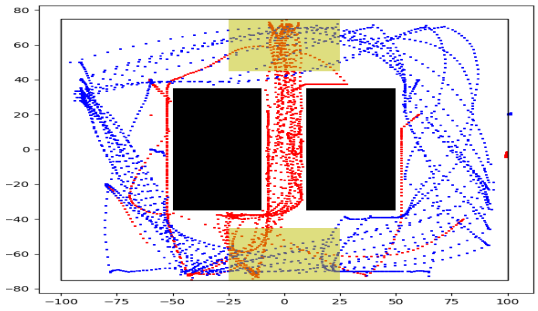

図 8: C100-振る舞い.

4.5 実験結果：学習速度の比較

本節では条件別に学習を行った際の学習収束時間について比較する．

4.5.1 C0

条件 C0 の学習過程である，獲得報酬遷移のグラフを図 9 と図 10 に示す．

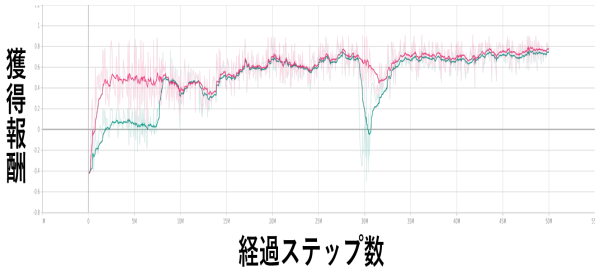

図 9: C0-報酬遷移 1 回目.

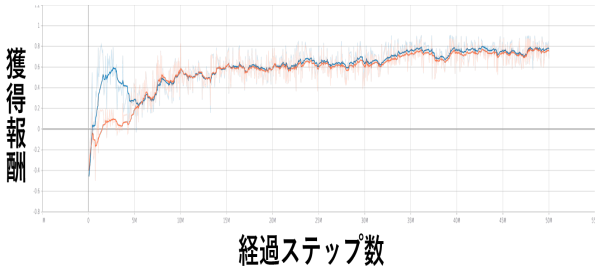

図 10: C0-報酬遷移 2 回目.

4.5.2 C25

条件 C25 の学習過程である，獲得報酬遷移のグラフを図 11 と図 12 に示す．

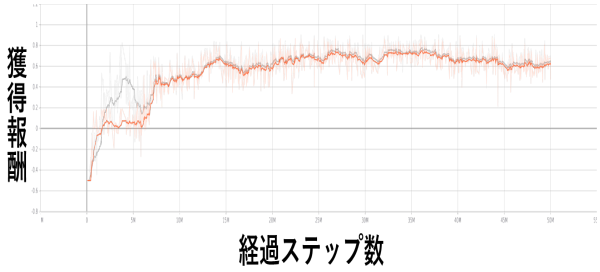

図 11: C25-報酬遷移 1 回目．

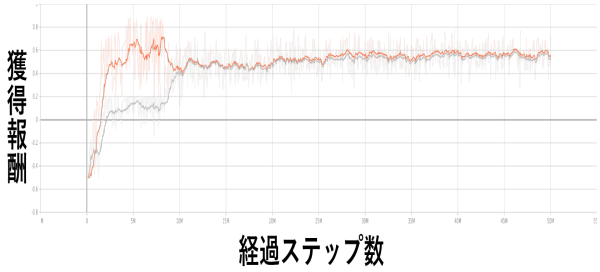

図 12: C25-報酬遷移 2 回目．

4.5.3 C50

条件 C50 の学習過程である，獲得報酬遷移のグラフを図 13 と図 14 に示す．

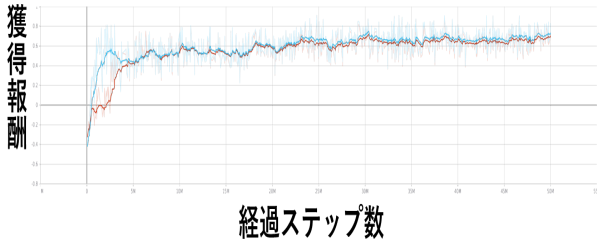

図 13: C50-報酬遷移 1 回目．

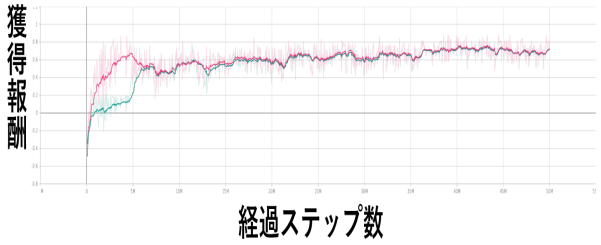

図 14: C50-報酬遷移 2 回目．

4.5.4 C100

条件 C100 の学習過程である，獲得報酬遷移のグラフを図 15 と図 16 に示す．

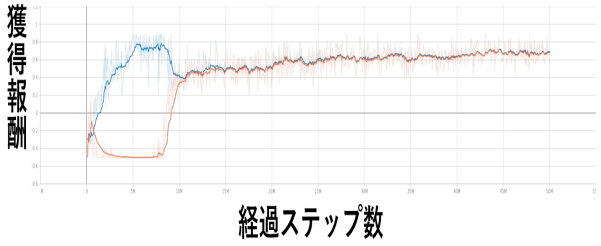

図 15: C100-報酬遷移 1 回目．

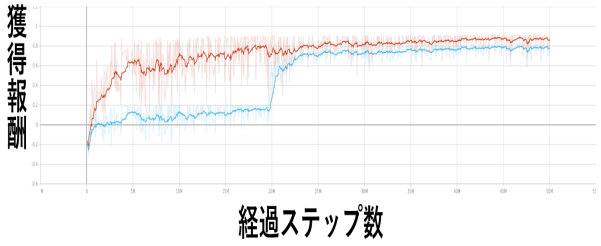

図 16: C100-報酬遷移 2 回目．

4.5.5 学習収束時間の比較

本実験では他者情報を参照する範囲を条件として，強化学習を行った．各条件で探索する状態変数を表 7 に示す．表 7 のように C0 の探索範囲が最小で，C100 の探索範囲が最大となる．条件ごとの学習収束までにかかった step 数を表 8 にまとめた．ここでの学習収束の定義は，学習の獲得報酬遷移をもとに指数移動平均 (tensorboard の smoothing 機能) の係数を 0.9 にしたとき，獲得報酬が 0.6 以上に達した時と定義している．

表 7: 条件別探索範囲

条件名	他エージェントとの相対位置の範囲
C0	X 座標，Z 座標 [0,0]
C25	X 座標， Z 座標 [-25,25]
C50	X 座標， Z 座標 [-50,50]
C100	X 座標，Z 座標 [-100,100]

表 8: 条件別学習の収束までの step 数

条件名	収束ステップ数 一回目	二回目
C0	16.97M[step]	15.02M[step]
C25	12.95M[step]	***
C50	18.15M[step]	15.7M[step]
C100	21.49M[step]	21.97M[step]

*** 報酬が基準値に満たなかった

5 考察

5.1 学習過程

本節では条件別の学習過程について考察を行う．4 条件の学習過程を報酬遷移を示している図 9，11，13，15 を見ると，それぞれの条件に学習の段階があることがわかる．学習の段階としては以下の 3 段階に分けられる．

(1) 初期のランダム方策による行動をとっている段階

(2) 一方のエージェントが多く報酬を得れるようになり，課題をクリアできているが，もう一方のエージェントは探索している段階

(3) 両方のエージェントが課題をクリアできるようになり，精度を高めるための学習を行っている段階

図 15 に示している条件 C100 の学習過程を見ると，1M ステップまでは第 1 段階のランダム行動による探索を行っており，その後一方のエージェントが報酬を得れるようになり，5M ステップ付近では課題をクリアできる方策を見つけている．もう一方のエージェントは報酬を得ておらず，9M ステップ付近まで探索を行っており，10M ステップを超えたあたりで報酬を得られるようになり，双方とも課題をクリアできる方策を見つけ安定して報酬を得ている．他の 3 条件では C100 ほど 2 体のエージェントの学習に差は見られないが，3 段階の学習段階に分かれている．これはそれぞれの学習において一方のエージェントが先に最適化を行うことでリーダーとなり，もう一方のエージェントはフォロワーとなり，そのリーダーに導かれる形で最適化を行ったと考えられる．以上のことから，複数体で環境を共有し，最適化を行っていくとき，学習度の異なるエージェントを用意すればリーダーフォロワー関係が生まれ，秩序の形成に寄与できると考えられる．

5.2 条件別学習による秩序の形成

本節では条件別の学習によって秩序が形成されたか，またどのような秩序が形成されたかについて考察を行う．学習後のエージェントの振る舞いをプロットしている図 5，6，7，8 を見るとすべての条件で 2 体のエージェントが別々のルートを走行していることがわかる．エージェントの走行ルートを条件別にまとめたものを表 9 に示す．

表 9: 条件別ルート選択

条件名	小さなエージェント	大きなエージェント
C0	サイド（右側）	真ん中
C25	真ん中	サイド（右側）
C50	サイド（一周）	真ん中
C100	真ん中	サイド（一周）

表 9 を見ると，一方のエージェントが真ん中の狭い道路を走行し，もう一方がサイドの広い道路を走行するといったルールがボトムアップに創発されている．秩序とはあらかじめ存在するルールとは別のボトムアップに創発した規則と定義しており，実験結果は表 9 のように 2 体のエージェント間でルートを分けて走行するといった規則をボトムアップに創発しているため，すべての条件において秩序の形成が行われたと考えられる．参照範囲を狭くし，他者の情報を限定した場合でも秩序の創発が観測できたことから，モデルを簡単化しても高い環境適応性を示せることが示唆された．今回創発された秩序は各エージェントが別々の経路を選択するといった単純な規則であった．これは 2 体のエージェントという最小個体数で環境を共有した学習であるため複雑なインタラクションが起こらなかったと考えられる．今後はエージェントの数と個性の要素を増やし，エージェントを多様化することで，創発される秩序にどのような変化があるか検証する．

5.3 情報を限定したときの学習速度の違い

本節では表 7 に示している探索範囲と表 8 に示している学習収束時間から考察を行う．強化学習では状態変数が増えたり，状態変数の範囲が大きくなると学習にかかる時間も増え，モデルの複雑さも増す．表 7 を見てもわかる通り，条件 C0 のとき，他者の情報を認知しないため探索範囲は少なく，学習時間も減少すると予測できる．また，条件 C25，C50，C100 と参照範囲を

増やしていくと探索する範囲は広がってくるため，学習にかかる時間も増えると予測していた．表 8 に示している学習収束時間を見ると，条件 C100 だけ 21Mstep を超え，収束までに一番時間がかかっていたため予測通りの結果となった．しかし，条件 C25 が一番早く収束しており，条件 C0 と条件 C50 には大きな差は見られなかった．この要因としてはエージェント数が少なく，最適化が簡単であったことが考えられる．2 体のエージェントで学習を行ったとき，状態変数の次元数は 9 であるが，エージェントを 10 体に増やすと状態変数の次元数は 36 に増加し，他者の情報を探索する範囲は指数状的に増加する．したがって，エージェント数を増やし同条件で学習すると，学習収束までにかかる時間に差がでてくると考えられる．

6　まとめと今後の計画

本研究では，荷運び課題を題材として，マルチエージェントに強化学習させることで秩序の形成を行わせた．他者情報の参照範囲を条件とした学習の結果，すべての条件においてエージェントのルート選択に秩序だった振る舞いが観測され，他者情報を限定しても秩序だった振る舞いを獲得できた．この結果は他者と環境を共有するとき，他者と直接インタラクションを行うのではなく環境とのインタラクションにより，秩序の形成が可能であることが示唆された．しかし，実際に行動するエージェント数の数が少なかったため，情報の限定し学習をしても学習速度に大きな差は見られなかった．今後はエージェント数を増やし，多種多様な行動主体を用意して学習を行ったときに，異なった秩序を創発するのか検証し，どのような過程で秩序が形成されたか解明する必要がある．また，本稿では学習後のモデルについて報酬が規定値を超えているか否かで評価を行っており，モデルの精度については詳細な議論をしていない．したがって，今後はモデルの精度について詳しく評価を行ってい，どのくらい最適化が行われたかについて条件別に比較する．

人間社会で創発される秩序について理解するためには，荷運び課題のような，環境と目的を共有し，多種多様な行動主体間でインタラクションが行われる環境の中で，どのように秩序が形成され，秩序形成に何が必要であるか，といったことを検証することが重要である．荷運び課題のように最適化を行うのに行動主体それぞれのインタラクションを考える必要がある複雑系の環境において，他者情報を限定した場合でも秩序を形成し，高い環境適応性を示すことができれば，環境とインタラクションを行い，少ないリソースのもとで他者の振る舞いに応じた行動をとれる，エージェントモデルの構築が期待できる．

参考文献

[1] 倉本, 安田, 山本, 水口, 辻野:対話エージェントへの「個性」の付与: 意思決定支援システムに対する影響; 情報処理学会　インタラクション 2012,pp.223-228 (2012).

[2] 岡, 森田, 大本:インタラクションを持続させる個人特性-システム化と共感に注目した検討;HAI シンポジウム (2018)

[3] K.Ozawa, T.Wakita, C.Miyajima, K.Itou, and. K.Takeda: Modeling of Individualities in Driving Through Spectral Analysis of Behavioral Signals; Institute of Electrical and Electronics Engineers, pp.851-854 (2005).

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

[7] 保田, 大倉: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).

[8] 大倉: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 (2011).

[9] 山田: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.17-00288 (2018).

[10] T.Parsons, 稲上毅:社会的行為の構造（第５分冊）; 木鐸社刊 (1989).

[11] Amy Greenwald, and. Keith Hall: Correlated-Q Learning; AAAI, pp.84-89 (2002).

[12] Yingfeng Cai, Shaoqing Yang, Hai Wang, Chenglong Teng, and. Long Chen: A Decision Control Method for Autonomous Driving Based on Multi-Task Reinforcement Learning; IEEE Access2021, pp.154553-154562 (2021).

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

[14] Uwano, F. and Takadama, K.: Utilizing Observed Information for No-Communication Multi-Agent Reinforcement Learning toward Cooperation in Dynamic Environment, SICE Journal of Control, Measurement, and System Integration, Vol. 12, No. 5, pp. 199-208 (2019)